Universidad Autónoma de Madrid

Facultad de Ciencias

Departamento de Biología Molecular

Analysis of the interaction between chromosomal replication and transposition mediated by sliding clamps

Doctoral Thesis

Héctor Díaz Maldonado

Thesis supervisor: Dr. Francisco J. López de Saro

Madrid, 2016

Doctoral thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

.

The research here presented was carried out at the Centro de Astrobiología (Instituto Nacional de Técnica Aeroespacial - Consejo Superior de Investigaciones Científicas) under the supervision of Dr. Francisco J. López de Saro and was funded by the Subdirección General de Proyectos de Investigación of the Spanish Ministry for Economy and Competitiveness Grants No. CGL2010-17384.

Héctor Díaz Maldonado was supported by a FPI fellowship from the Spanish Ministry for Economy and Competitiveness.

INDEX

Index of figures

Index of tables

Abbreviations

Å	Ångström
Ac	*Acidiphilium sp.*
Af	*Acidithiobacillus ferrivorans*
BLAST	Basic local alignment search tool
Bp	Base pair
BSA	Bovine serum albumin
C-	Carboxi
cDNA	Complementary DNA
DNA	Deoxyribonucleic acid
dNTPs	deoxynucleotides
e.g.	exempli gratia (for example)
Ec	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
et al.	et alii (and others)
GST	Glutathione S-transferase
h	Hour
HGT	Horizontal gene transfer
i.e.	id est (that is)
IE	Inside end
IPTG	Isopropyl β-D-1-thiogalactopyranoside
IR	Inverted repeats
IS	Insertion sequence
Kb	Kilobase pairs
KDa	Kilodalton
Lf	*Leptospirillum ferooxidans*
mA	Milliampere
Mb	*Methanosarcina barkeri*
min	Minute

mg	Miligrame
ml	Milliliter
mM	Millimolar
mRNA	Messenger RNA
mV	Millivolts
N-	Amino
ng	Nanogram
nm	Nanometers
nM	Nanomolar
OD	Optical density
OE	Outside end
ORF	Open reading frame
PAGE	polyacrylamide gel electrophoresis
PBS	Phosphate buffered saline
PCNA	Proliferating cell nuclear antigen
pmol	Picomol
RNA	Ribonucleic acid
SB	Supernatant
SD	Standard deviation
SDS	Sodium dodecyl sulfate
Tn	Transposon
Tnp	Transposase
UV	Ultraviolet
w/v	Weight/Volume
Wt	Wild-type
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
β	β subunit of the DNA polymerase III
μg	Microgram
μl	Microliter
μM	Micromolar

Abstract

Insertion sequences (ISs) are small mobile genetic elements widely distributed in prokaryotes. They often encode only one enzyme, the transposase, required for their own transposition. ISs are promiscuous elements that can proliferate within genomes, where they play a key role in genome evolution by promoting chromosomal rearrangements and genetic flow. Furthermore, ISs have the ability to cross species barriers and transpose actively in new hosts, which also makes ISs essential players in the process of horizontal gene transfer. Although highly autonomous, ISs activity is linked to and can be regulated by various host processes, especially chromosomal replication; however no general mechanism had been proposed connecting replication with transposition.

In this thesis we investigated the interplay between transposases and host replication factors. First, we performed a survey of orientation patterns of IS in fully-sequenced bacterial chromosomes. We found that a significant fraction of IS families present a consistent and family-specific orientation bias with respect of the movement of the replication fork, especially in Firmicutes. Then, we found that the transposases of up to ten different IS families with different transposition pathways interact with *E. coli* β sliding clamp, an essential replication factor. Additionally, we demonstrated that purified transposase of Tn*5* also interact with β sliding clamp. Moreover, we studied to what extent the interaction limits or favors the ability of ISs to colonize a chromosome from a phylogenetically-distant organism. We describe the proliferation of a member of the IS*1634* family in a long-term culture of *Acidiphilium sp.* We found that the *Acidiphilium* IS*1634* transposase binds to β sliding clamp of *Acidiphilium, Leptospirillum* and *E. coli*. Further, we also demonstrated that *Acidiphilium* IS*1634* transposase binds to the archaeal sliding clamp (PCNA) from *Methanosarcina*, and that the transposase encoded by *Methanosarcina* IS*1634* binds *Acidiphilium* β. Finally, we demonstrated that strengthening the interaction between β and the transposase results in an increased transposition rate *in vivo*.

Our results strongly suggest that transposase interaction with sliding clamps is a widespread mechanism that allows ISs integration with host chromosomal replication. Interaction with β and asymmetries in β distribution in the replication fork could explain the observed strong orientation bias found in some IS families in Firmicutes. Sliding clamps may represent a universal and highly conserved platform for ISs dispersal between species. The strength of the interaction could determine the potential of ISs to be mobilized in bacterial populations and also their ability to proliferate within chromosome.

Resumen

Las secuencias de inserción (SI) son pequeños elementos genéticos móviles ampliamente distribuidos en procariotas. Habitualmente codifican para una solo enzima, la transposasa, requerida para su propia transposición. Las SI son elementos promiscuos que puede proliferar en los cromosomas, donde juegan un papel clave en la evolución de los genomas promoviendo la reorganización cromosómica y el flujo genético. Además, las SI tienen la habilidad para cruzar la barrera inter-especie y transponerse activamente en nuevos huéspedes, lo que también las convierte en actores esenciales en procesos de transferencia génica horizontal. Aunque son altamente autónomas, la actividad de las SI está ligada y puede ser regulada por varios procesos del hospedador, especialmente la replicación del cromosoma; sin embargo no se ha propuesto ningún mecanismo general conectando la replicación con la transposición.

En esta tesis investigamos la interacción entre transposasas y factores de replicación del hospedador. Inicialmente, realizamos un estudio de los patrones de orientación de las SI en cromosomas bacterianos completamente secuenciados. Encontramos que una fracción significativa de familias de SI presentan un sesgo de orientación consistente y específico de la familia, con respecto al movimiento de la horquilla de replicación, especialmente en Firmicutes. Además hallamos que la transposasa de hasta 10 familias distintas de SI con diferentes mecanismos de transposición, interaccionan con β *sliding clamp* de *E. coli*, un factor esencial de la replicación. Asimismo, demostramos que la transposasa purificada de Tn*5* también interacciona con β *sliding clamp*. Además estudiamos hasta qué punto esta interacción limita o favorece la habilidad de las SI para colonizar cromosomas de organismos distantes filogenéticamente. Describimos la proliferación de un miembro de la familia IS*1634* en un cultivo de larga duración de *Acidiphilium sp.* y demostramos que la transposasa de IS*1634* de *Acidiphilium*, interacciona con β *sliding clamp* de *Acidiphilium, Leptospirillum* y *E. coli*. Más aún, demostramos que la transposasa de IS*1634* de *Acidiphilium* también interacciona con el *sliding clamp* de la arquea (PCNA) *Methanosarcina*, y que la transposasa codificada por IS*1634* de *Methanosarcina* interacciona con β de *Acidiphilium*. Finalmente, demostramos que fortaleciendo la interacción entre β y la transposasa resulta en un incremento en la tasa de transposición *in vivo*.

Nuestros resultados sugieren consistentemente que la interacción entre transposasa y *sliding clamp* es un mecanismo ampliamente distribuido que permite a las SI integrarse con la replicación de los cromosomas hospedadores. La interacción con β y las asimetrías en la distribución de β en la horquilla de replicación, podrían explicar los fuertes sesgos en la

orientación de ciertas familias de SI en Firmicutes. Los *sliding clamps* pueden representar una plataforma universal y altamente conservada para la dispersión de las SI entre especies. La afinidad de la interacción puede determinar el potencial de las SI para movilizarse en poblaciones bacterianas y también su habilidad para proliferar dentro de los propios cromosomas.

Introduction

1.1. The genetic transposable elements

"It takes all the running you can do, to keep in the same place", the Red Queen told Alice in *Through the Looking-Glass, and what Alice found there*. This passage inspired the name of the well-known evolutionary theory of the *Red Queen* (Van Valen 1974): genomes face a constant evolutionary run, developing new traits, to successfully adapt to an ever-changing environment or the appearance of new competitors. Genetic transposable elements are major players in the creation of genetic variability that allows genomes to be on the run. Transposable elements were first described by the Nobel laureate Barbara McClintock in maize (McClintock 1950), but virtually they can be found in any branch of the tree of life. In fact, proteins annotated in databases as transposases or transposase-related proteins are the most abundant and most distributed functional class, both in prokaryotes and eukaryotes (Aziz *et al.*, 2010).

Transposition implies the movement of a discrete segment of DNA (i.e., the transposable element), from a donor site to a target location elsewhere in the genome. Transposable elements are divided into two principal categories: DNA transposons that move via DNA (class II), and retrotransposons that move using an RNA intermediate (class I). They can also be classified as autonomous o non-autonomous elements, depending on whether they encode the enzyme required for the transposition reaction, the transposase (Craig *et al.*, 2002). Insertion sequences (IS) are commonly referred to as the simplest DNA transposable element, and they are ubiquitous in bacterial genomes. ISs are usually between 0.5 and 2.5 kb in length and often have just a single gene that encoding the transposase enzyme. The transposase gene is usually flanked by terminal inverted repeats. Some ISs also have short direct repeats generated in the target DNA as a consequence of the insertion process (Mahillon and Chandler 1998). ISs generally encode no functions other than those involved in their mobility and propagation, but in some instances they carry passenger genes, that confer an advantage to the host, such as antibiotic resistance genes. ISs have been classified into families and subgroups based on diverse characteristics and, mainly, in the sequence of the transposase enzyme (Table 1.1; Mahillon and Chandler 1998).

1.2. Transposase diversity and biochemistry

Transposases employ various enzymatic mechanisms to catalyze the DNA breaks, strand transfer, and strand joining reactions involved in the process of transposition (Curcio and Derbyshire 2003). According to the catalytic mechanism, transposases can be grouped in

the following families: DDE-transposases, HUH-transposases, Y-transposases and S-transposases (Hickman and Dyda 2014). Additionally, some transposons, termed retrotransposons, duplicate via RNA intermediates that are reverse transcribed and inserted elsewhere in the genome. Retrotransposon enzymes usually combine a reverse transcriptase and endonuclease activities (Craig *et al.*, 2002). They are usually present in eukaryotic genomes and are classified by the presence or not of Long Terminal Repeats (LTR), being the non-LTR group the most abundant in the human genome, accounting for one-third of its length (de Koning *et al.*, 2011). Non-LTR retrotransposons includes Long Interspersed Element-1 (LINE-1) and Short Interspersed Elements (Alu, and SVA elements).

Family	Subgroups	Catalytic residues
IS*1*	Single ORF	DDE
	IS*Mhu11*	
IS*1592*	IS*Pna2*	DDNK
	IS*H4*	
	IS*1016*	DDEK
	IS*1595*	DDNK + ER4R7
	IS*Nwi1*	DDEK+ ER4
	IS*Nha5*	DDER/K
IS3	IS*150*	DDE
	IS*407*	
	IS*51*	
	IS*3*	
	IS*2*	
IS481		DDE
IS*4*	IS*10*	DDE
	IS*50*	
	IS*Pepr1*	
	IS*4*	
	IS*4Sa*	
	IS*H8*	
	IS*231*	
IS*701*	IS*Aba11*	DDE
IS*H3*		DDE
IS*1634*		DDE
IS*5*	IS*903*	DDE
	IS*L2*	
	IS*H1*	
	IS*5*	
	IS*1031*	
	IS*427*	

Family	Subgroups	Catalytic residues
IS*1182*		DDE
IS*6*		DDE
IS*21*		DDE
IS*30*		DDE
IS*66*	IS*Bst12*	DDE
IS*256*	IS*1249*	DDE
	IS*C1250*	
IS*H6*		DDE
IS*Lre2*		DDE
IS*Kra4*		DDE
IS*630*		DDE
IS*982*		DDE
IS*1380*		DDE
IS*As1*		
IS*L3*		
Tn*3*		DDE
IS*Azo13*		
IS*110*	IS*1111*	
IS*91*		HUH/Y2
IS*200*/IS*605*	IS*200*	HUH/Y1
	IS*605*	HUH/Y1
	IS*1341*	
IS*607*		Serine
ISNCY	IS*892*	
	IS*Lbi1*	
	IS*Mae2*	
	IS*Plu15*	
	IS*A1214*	
	IS*C1217*	
	IS*M1*	

Table 1.1 Classification of bacterial IS families.
Subgroups that conform each IS family and their respective catalytic amino acids are shown. Modified from Siguier *et al.*, 2015.

1.2.1. The DDE transposase superfamily

The best characterized and most abundant transposase group is the DDE-transposase, so-called after the non-contiguous amino acid triad that commonly conform the active site (Asp, Asp, Glu). Most transposases of prokaryotic insertion sequences, the eukaryotic Mariner elements, and the retroviral integrases like the avian sarcoma virus integrase (Bujacz *et al.*, 1995) or the HIV-1 integrase (Dyda *et al.,* 1994) belong to the DDE-transposase group. The enzyme RAG1 which catalyzes a transposition-like chemistry in the V(D)J recombination system which generates immunoglobulin diversity (van Gent *et al.,*1996; Hiom *et al.,*1998) is also a member of the DDE family.

Although members of this group are divergent in sequence, the structure and location of the catalytic domain is conserved (Hickman *et al.*, 2010). The DDE active site is an RNaseH-like catalytic domain, where the three amino acid residues coordinate two divalent metal ions (Lovell *et al.*, 2002). The active site can cleave multiple DNA strands by using reiterative steps of hydrolysis and trans-esterification. First, a nucleophilic attack by an activated water molecule in the transposon ends breaks the phosphodiester DNA bond generating a 3´OH transposon intermediate (Mizuuchi and Baker 2002). Then, in a trans-esterification reaction, the exposed 3´OH cleaves the target DNA and simultaneously joins the transferred and target strands (Kennedy *et al.*, 2000). The repair by host enzymes of any gap left behind generates the characteristic directed repeats for each transposon of this group. However, members of this family have also notable differences regarding the chemistry of the intermediate form.

For Tn*5* or Tn*10* the first nucleophilic attack is on one transposon strand (i.e., transferred strand) which subsequently generates a transposon hairpin intermediate (Kennedy *et al.*, 1998; Bhasin *et al.*, 1999). In contrast, for Mos1, a eukaryotic Mariner element, the initial nucleophilic attack is on the non-transferred strand (Dawson and Finnegan 2003). All of them transpose via a *cut-out, paste-in* mechanism (Fig. 1.1 A).

On the other hand, bacteriophage Mu or transposons related to Tn*3* family use a mechanism whereby transposition is coupled to DNA replication (Nicolas *et al.*, 2015). Only one side of each strand of the transposon is cleaved generating a 3´-OH. Then, the

transposon and its flanking DNA attack and join the target site. The other strand is synthesized by the DNA replication machinery (Curcio and Derbyshire 2003) (Fig. 1.1 C).

Yet another replicative transposition pathway involves IS*3* like elements such as IS*911* (Hickman and Dyda 2014). The first nucleophilic attack by water on the transposon end generates a 3´-OH group that is then used to attack the same strand at the opposite end. Then, the replication machinery of the host generates an excised transposon circle intermediate which repairs the donor DNA gap. The intermediate is inserted into the target site (Duval-Valentin *et al.*, 2004). This mechanism is named as *copy-out, paste-in* (Fig. 1.1 B).

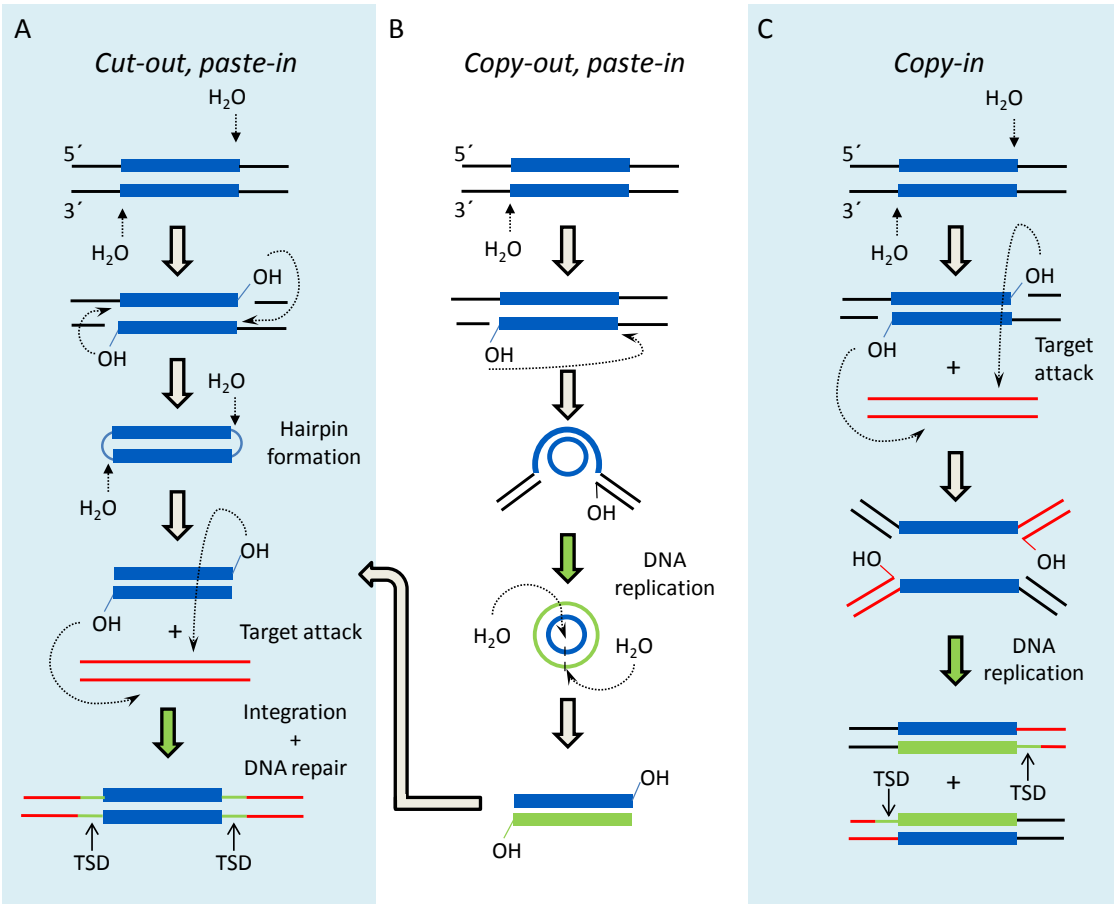


Figure 1.1. Mechanisms of DDE-transposases. A) Cut -out, paste-in mechanism followed by most of transposases of IS*4* family, like Tn*5* or Tn*10*. B) *copy-out, paste-in* followed by IS*3* or IS*911*. C) Co-integration mechanism of Tn3. Blue rectangles represent both strands of the transposon. Black lines the flanking DNA. Red lines are the target DNA. Green lines or green rectangle are replicated or repair DNA by the host replication machinery. Nucleophilic attacks are denoted by dotted lines. Target site duplication (TSD) is also shown.

1.2.2. The HUH transposase superfamily

HUH-transposases are the second major group of transposases. They transpose via a ssDNA intermediate mechanism. They have a nuclease domain that coordinates a single

metal ion in a motif formed by two histidine residues (H) separated by an hydrophobic amino acid (U) (Dyda *et al.*, 2012). In HUH-transposases a nucleophilic attack is carried out by the -OH group of either one tyrosine (Y1-transposases) (Ronning *et al.*, 2005) or two close tyrosines (Y2-transposases) located in the catalytic domain. The cleavage product is a single strand DNA with a 3´-OH group and a 5´- phosphotyrosine linkage (Chandler *et al.*, 2013). Then, the 3´-OH group attacks the 5´- phosphotyrosine, creating a single-stranded circular transposon intermediate (Hickman and Dyda 2014). The integration mechanism does not generate target site duplication. Best characterized members of this group are the IS*200* family (Y1-transposases) and IS*91* (Y2-transposases).

The transposition of IS*608*, a member of the IS*200*/IS*605* family, has been studied in detail. It requires a ssDNA substrate and a dimeric enzyme catalyzes the strand-specific cleavage and transfer. The intermediate is a circle transposon that is integrated in a ssDNA target (Barabas *et al.*, 2008; He *et al.*, 2013). Interestingly, IS*608* insertions exhibit an orientation preference for the lagging-strand template and insertion can be specifically directed to stalled replication forks (Ton-Hoang *et al.*, 2010).

In the other hand, IS*91* (Y2-transposase) generates a circular intermediate that has been proposed to follow a transposition process that resemble a rolling circle mechanism (Garcillán-Barcia *et al.*, 2001).

1.3. IS impact in genomes: IS expansion and genome evolution

The dynamics and capacity for proliferation of ISs have fed vigorous debates about their impact in genomes. For example, IS have been considered 'genomic diseases' when transposons expand and eventually inactivate the genomes that carry them (Wagner 2009), although possibly the host organism could actively impose regulatory mechanisms to control IS proliferation (Zeh *et al.*, 2009). Other hypothesis stresses the fact that IS abundance could also be neutrally regulated by bacterial population size (Linch and Conery, 2003). It has been proposed that at large evolutionary scales IS expansion is controlled by deletions rather than purifying selection. Under this view a neutral dynamic equilibrium state is established between IS loss by deletions and IS proliferation by duplication and HGT. The components of this neutral equilibrium could be transitorily perturbed resulting in a rapid expansion of IS population (Touchon and Rocha 2007; Iranzo *et al.*, 2014).

Genomes often contain significant amounts of truncated or partial transposase sequences. These truncated transposases are molecular fossils that represent IS expansion and extinction cycles over evolutionary time, and can be used to track the evolutionary past

of the host organism (Cerveau *et al.*, 2011; Wagner 2006). Functionally, truncated IS sequences remain in the chromosome are genetic scars of previous infections and could also play subtle roles in inhibiting transposition of elements of the same family (Gueguen *et al.*, 2006). The analysis of the intragenomic variation of IS reveals low sequence diversity of a given IS family within a chromosome. This suggest that active ISs in genomes usually are evolutionary young and have been recently acquired in distinct waves of IS invasion (Wagner 2006). This observation reinforces the idea that IS activity does not occur at a constant rate, but rather in abrupt blooms or bursts (Olivier and Greene 2009).

IS abundance and distribution are also a matter of discussion. ISs are widespread among prokaryotes and represent around the 3% of prokaryote genomes (Siguier *et al.*, 2006). However ISs are highly skewed and patchy distributed between species or even between individuals. In some organisms ISs account for a large proportion of the genome, like in Wolbachia wRi where ISs represent up to 11% of the genome (Cerveau *et al.*, 2011). However, in other species there is not a single IS, such is the case of *Bacillus subtilis* (Kunst *et al.*, 1997). Reasons of IS abundance in organisms are not clear, although it has been proposed that bigger genome size, higher frequency of horizontal gene transfer and ecological associations correlates with higher IS abundance (Touchon and Rocha 2007).

IS proliferation has a direct impact in chromosome structure and genome evolution. IS expansion can induce gene duplication and deletion, and are potentially mutagenic by disrupting host coding sequences or their regulatory regions. Besides, many ISs have their own promoters, so IS insertions can modify the expression levels of neighboring genes (Nagy and Chandler 2004). Another key contribution of transposable elements is as a source of raw material in the evolution of new genes by means of molecular domestication, as for example the evolution of V(D)J system (Kapitonov and Jurka 2005; Cordaux *et al.*, 2006). However the profoundest effect of IS expansion is the promotion of chromosomal homologous recombination and genomic rearrangements when a given IS is in high copy number (Rocha 2003). After IS massive proliferation, ISs will have a tendency to undergo deletion with adjacent DNA sequences in the absence of direct selection, which has an effect in genome size reduction in isolated bacteria or with host-restricted lifestyles, like endosymbionts (Siguier *et al.*, 2014). Collectively, transposable elements shape genome structure and generate the genetic variability in populations required for rapid diversification of taxa.

ISs do not commonly confer a direct a selective advantage to the host. Thus, IS persistence in an organism implies that they should be constantly duplicating themselves or

been imported in the genome by HGT to avoid removal by genetic drift or deletion (Hooper *et al.*, 2009; Iranzo *et al.*, 2014). Thereby, it could be speculated that when imported by HGT, IS have a short time window to integrate in the genome before they are inactivated by mutations (Le Rouzic and Capy 2005; Bichsel *et al.*, 2010). However, the effective invasion of a new host relies on the compatibility of the IS with the new molecular environment. This implies that the transposase gene has to be transcribed and translated, and any interactions of the transposase with host functions ought to be maintained.

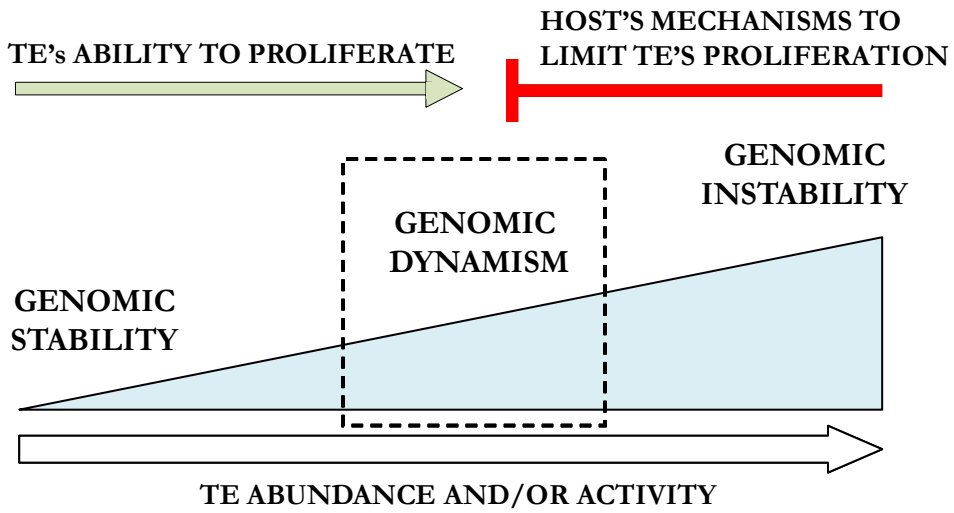


Figure 1.2. Hypothesized relationship between transposable element activity, host regulatory mechanisms and genomic stability.
An explosive and uncontrolled expansion of transposable elements (TE) could lead to the extinction of bacterial lineages (Wagner 2006). Therefore, a subtle equilibrium between IS proliferation and IS control must be established to reach an optimal state of genome dynamism to allow adaptation and evolution. Modified from Olivier and Greene 2009.

1.4. IS impact in genomes: horizontal gene transfer

Comparative genomics of bacteria populations have revealed the existence of a high genetic diversity among individuals of a single species. For instance, a study of 20 strains of *E. coli* highlighted that up of 90% of their pan-genome are accessory and strain-specific genes (Touchon *et al.*, 2009). The variation in genetic content reflects how genetic information is continuously exchanged between bacterial communities, mobilized by diverse mechanisms and recombined at high rates. One of the major driving forces of bacterial adaptation and evolution is horizontal gene transfer (HGT) between organisms

not necessarily phylogenetically related (Toussaint and Chandler 2012). The three main processes implicated in HGT are natural transformation by DNA uptake from the surrounding environment, transduction mediated by bacteriophages and conjugation facilitated by plasmids through contacts between bacteria (Thomas and Nielsen 2005).

Transposable elements have a key role in the rapid acquisition by HGT of new mechanisms that allow bacterial adaptation to environmental fluctuations or explore new ecological niches (Casacuberta and González 2013). While eukaryotes typically evolve via gene duplication and neofunctionalizations, prokaryotic organisms evolve in short evolutionary times by new gene acquisition via HGT.

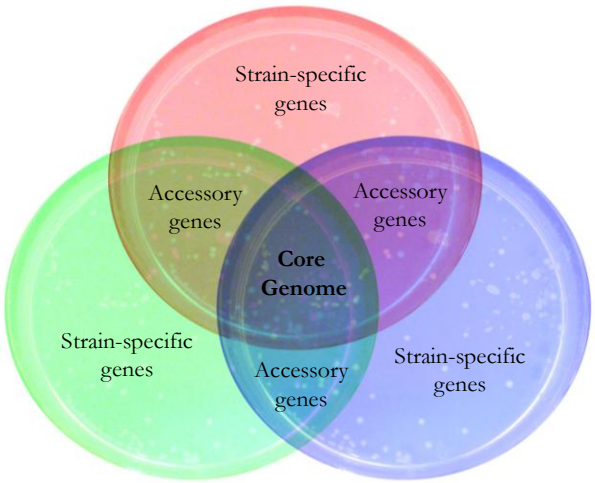


Figure 1.3. Venn diagram representing the pan-genome of three hypothetical strains.
The pan-genome is defined as the complete set of genes that can be present in a given species. It is composed of a core genome that encompasses all common genes to the diverse strains of the species. Additionally, organisms possess a variable and strain-specific genome that is the result of genetic exchange and recombination (Feil 2004; Medini *et al.*, 2005). ISs critically contribute to shape genomes and promote diversification and speciation.

The characteristic feature of transposons and ISs is that they can move highly autonomously between replicons, either within the bacterial chromosome itself or between phage or plasmids. ISs can be inserted in phage genomes or recombined with resident prophage IS elements and spread across phage bacterial hosts (Hendrix 2003). Moreover, plasmids are the most important vehicles for HGT in bacteria and critically contribute to host cell adaptability and fitness (Wozniak and Waldor 2010). Those that are able to self-transfer between different species usually have a larger size than other plasmids because they carry transfer-related genes, and they are also able to acquire transposons and insertion sequences while they infect different host genomes. An initial survey of plasmids detected that those bigger than a certain threshold (> 20Kb) tend to accumulate a higher percentage

of ISs than the host bacterial chromosome (Siguier *et al.*, 2006). For instance the *Shigella flexneri* virulence megaplasmid pw100 is almost 50% of its length composed of ISs and IS-related truncated sequences (Venkatesan *et al.*, 2001).

Additionally, there are some reported examples of preferential plasmid-targeting by transposons. Tn*7*, a transposon that encodes five different proteins (TnsA, B, C, D and E), has variable target pathways that are regulated by presence of either TnsD or TnsE in the transposase protein complex (Rogers *et al.*, 1986, Kubo and Craig 1990). TnsE is a DNA-binding protein that recognizes features of the discontinuous (lagging strand) synthesis. In the pathway directed by TnsABC+E, Tn*7* insertions take place preferentially in transmissible plasmids in an orientation specific manner during their replicative transfer between different bacteria (Peters and Craig 2001). Thereby, transposons and IS are relevant vehicles moving along the horizontal transfer highways.

Transposons and IS do not usually travel alone. They also facilitate the flow of other genes between replicores. They promote horizontal movement of genetic information either by directly carrying *passenger* genes or by facilitating homologous recombination between replicores when they are present in high copy number. This genetic transfer leads to the appearance of new bacterial traits and functions. It has been widely documented that transposons are closely bound to the acquisition and spread of antimicrobial resistance (Levy and Marshall 2004; van Hoek *et al.*, 2011). This phenomenon represents a global health concern (WHO 2014). HGT of antibiotic resistance is so widespread that resistance genes associated with transposons are even found in the microbiome of human populations that have never been exposed to antibiotics (Clemente *et al.*, 2015). In this niche there is no selective pressure, so antibiotics resistant genes were constantly imported by HGT to guarantee their persistence in the microbial population. Well-known transposons that carry multi-antibiotic resistance genes are Tn*5,* Tn*10* or Tn*21* (Reznikoff 1993; Kleckner 1989; Martinez and de la Cruz 1990). IS are also related to the dissemination of virulence genes, like IS*91* family elements that are commonly found surrounding toxic genes in pathogenic enterobacterial strains (Garcillán-Barcia and de la Cruz 2002).

IS elements have often been observed flanking recently-incorporated foreign DNA in genomic islands or in smaller gene clusters of environmental insolated bacteria, possibly facilitating the integration in the chromosome of new acquired traits (Juhas *et al.*, 2009). Genomic islands are segments of the chromosome which have been transferred via HGT in the recent evolutionary past, as revealed by their anomalous nucleotide composition

(Dobrindt *et al.*, 2004). For example genomic comparative analysis of two co-isolated strains of the halophylic bacteroidetes *Salinibacter ruber*, revealed that most divergent orthologous genes and strain-specific genes are concentrated in IS-enriched genomic islands (Mongodin *et al.*, 2005; Peña *et al.*, 2010).

The frequency of genes horizontally transferred and integrated in the chromosome decreases with phylogenetic distance (Williams *et al.*, 2012). This is likely due to the reduced efficiency of homologous recombination between distant organisms. On the other hand, a large-scale study of the distribution of transposases across taxons and habitats showed a significant correlation between habitat-sharing organisms and their transposases (Hooper *et al.*, 2009). This suggests that lateral transfer of transposases or genes is more likely to occur in organisms that share the same or overlapping habitats (Papke *et al.*, 2004). Although gene transfer is common between closely related bacteria, there are many reported examples where the transmission is established between phylogenetically-distant organisms. An extraordinary example of HGT has been demonstrated with the sequencing of the genome of the acidophilic red algae *Galdieria sulphuraria* (Schönknecht *et al.*, 2013). Comparative genomics have shown that adaptation to the acidic environment, heavy-metal resistance, and metabolic versatility can be directly attributed to at least 75 separate gene acquisitions from Archaea and Bacteria. Interestingly, genes recruited by HGT into the algae's genome are specially enriched in those from extremophilic Bacteria, suggesting that HGT between distant related but co-habiting organisms could be more extended than it was though before (Schönknecht *et al.*, 2013). Hence, the study of integrative elements, like transposases, contained in horizontally transferred traits could be used to track their evolutionary history and origin.

1.5. Transposition regulation

Looking through the prism of evolutionary scale, transposable elements contribute to adaptation and diversity of microbial populations by promoting chromosomal rearrangements, regulating genes expression, or facilitating HGT, among other processes. Nevertheless, IS activity can generate chromosome instability and are potentially mutagenic when inserted within essential genes. Additionally, transposase activity can be potentially toxic and interfere with cellular processes. For instance Tn*5* transposase is cytotoxic when over-expressed (Weinreich *et al.*, 1994a). Thus, transposons and ISs might also be seen as deleterious elements that could damage host genome integrity when unregulated.

A subtle equilibrium must be established between IS or transposase activity to ensure IS stability in the chromosome (Fig. 1.2). Regulation could be imposed by the host or by auto-regulatory mechanisms of the IS itself. For example, new insertions of Tn*7* avoid targeting DNA which already contains copies of Tn*7*. This is a mechanism named 'target immunity' and in Tn*7* is driven by the interaction of TnsB and TnsC and the recognition of pre-existing copies of Tn*7* (Skelding *et al.*, 2003). Additionally, transposition could be regulated at various steps, such as transcription, translation and transposase stability and activity. Hereafter, we briefly describe some intrinsic control mechanisms of transposition (i.e., imposed by the transposable element itself).

First, it has been speculated that transposase transcription is regulated by weak endogenous promoters (Nagy and Chandler 2004). However, this mechanism can not be considered as a general rule to all IS and further study of each particular element is required. Promoters are usually located within the inverted repeats of the IS, and transposase derived proteins (e.g., truncated transposases) could bind to them down-regulating its expression. This is the case for IS*911* (Gueguen *et al.*, 2006) or IS*1* (Escoubas *et al.*, 1991). mRNA stability and the formation of secondary structures also regulates transposase transcription. For instance Tn*5* transposase transcripts from neighboring genes include various repeated sequences that create mRNA secondary structures that mask the ribosome binding site. On the other hand, endogenous transcripts do not form such structures (Krebs and Reznikoff 1986; Schulz and Reznikoff 1991). This effect prevents transcriptional read-through form surrounding genes. Repressing mRNA stem-loops are also formed in transcripts from own promoter in other transposases like in IS*200* (Beuzón *et al.*, 1999).

A second level of control is the translation regulation. An anti-sense RNA transcribed in opposite orientation than the codifying RNA of the IS*10* transposase, could sequester translation initiation sequence, preventing the synthesis of the transposase (Ma and Simons 1990). Translational frame shifting (Chandler and Fayet 1993) and translation termination are also involved in regulation.

A further regulation step concerns the transposase biochemistry, such as auto-inhibitory mechanisms, DNA target site recognition, transposase supra structure assembly (i.e., transpososomes) or transposase interaction with other host proteins. Below we cover in detail the mechanism of Tn*5*, focusing on its transposase activity as a study model of transposition and its regulation.

1.6. Tn*5*: an auto-regulated transposition paradigm

1.6.1 Tn*5* structure and transposition mechanism

Since it was initially discovered in the mid-seventies (Berg *et al.*, 1975), the extensive data available generated by genetic, biochemical and structural studies placed Tn*5* as a perfect model for understanding the biochemical mechanisms behind transposition. Tn*5* is a prokaryotic transposon that belongs to the IS*4* family (see table 1.1). It is composed of two IS*50* copies, IS*50*L and IS*50*R, which encompass three antibiotics resistance genes (Reznikoff 1993). The transposon is delimited by two inverted repeats of 19 bp (outside ends or OE), required for Tn*5* transposition (Johnson and Reznikoff 1983). Each IS*50* element is enclosed by an inside end (IE) and an outside end (Fig 1.4). Tn*5* transposase (Tnp) is a 476 aa enzyme translated from a gene located in IS*50*R and orientated inwards the transposon. Effective transposase translation occurs from promoters situated close to the right OE. mRNA read-through from flanking genes contains a symmetry element that generates an unproductive RNA secondary structure (Schulz and Reznikoff 1991). IS*50*R also codifies for an inhibitor protein (Inh) that has the same full length transposase sequence, but lacking the N-terminal 55 aa. The Inh promoter is different from that of Tnp, and read-through transcripts do not form secondary structures. Hence, read-through transcripts from neighboring genes could synthesize inhibitor protein but do not active transposase. In the other hand, IS*50L* differs from IS*50*R in a single amino acid that introduces a stop codon, so transcripts from IS*50L* generate a truncated and inactive transposase (Rothstein and Reznikoff 1981).

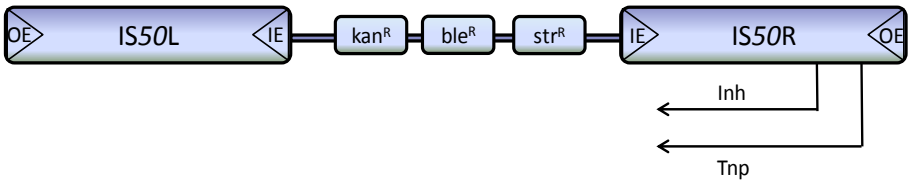


Figure 1.4. Transposon Tn*5* structure.
Two IS*50* copies embrace resistance genes for Kanamycin (kan), bleomycin (ble) and streptomycin (str) antibiotics. Outside ends (OE) and inside ends (IE) are represented. Inh and Tn*5* transposase (Tnp) are translated from IS*50*R.

Tn*5* transposes via a "copy-paste" mechanism. As other members of the IS*4* family, Tnp has a DDE active site that chelates two Mg^{+2} ions necessary for transposition (Mizuuchi and Baker 2002). The Tn*5* transposition mechanism can be studied at three

different levels. First, Tnp has to find and bind the DNA inverted repeats and form the synaptic complex. Further steps take place within the context of the synaptic complex. The second step involves three catalytic reactions in both transposon ends and results in DNA cleavage. A strand of the transferred DNA (i.e. transposon) is nicked by a nucleophilic attack of a water molecule that has been activated by Mg^{+2} ions at the Tnp active site. This attack on the phosphodiester backbone of a nucleotide in the junction between OE and flanking DNA, generates a 3´OH. Then, the 3´OH acts as a nucleophile and attacks the 5´ of a nucleotide in the opposite non-transferred strand (i.e., flanking DNA), generating an interstrand phosphodiester bond or hairpin intermediate (Bhasin *et al.*, 1999). The hairpin intermediate is resolved by a third nucleophilic attack of another H_2O molecule activated by Mg^{+2}. The product of this reaction is a blunt-end transposon released from the DNA donor backbone (Goryshin and Reznikoff 1998). The release of the transposon seems to be sequential, firstly by the break on one end followed by the catalysis at the second one (Steiniger *et al.*, 2006). These transposition reactions involving an intermediate are common for other transposases like Tn*10*, where the hairpin intermediate can also be found (Kennedy *et al.*, 1998).

Finally, in the strand transfer step, the post-cleavage synaptic complex captures the target DNA. Free 3´OH at both ends of the transposon attack the phosphodiester backbone of target DNA via a trans-esterification reaction (Mizuuchi *et al.*, 1991). A covalent bound is established between the 3´hydroxyl group of the transposon and the 5´phosphate groups of the target DNA, resulting in the integration of the transposon. Subsequently, a strand transfer reaction followed by the action of host repair enzymes generate the characteristic 9-bp duplication in the integration site (Reznikoff 2002).

1.6.2 Tn*5* transposase biochemistry and DNA binding

One of the key processes of Tn*5* transposition is repeat-sequences recognition by the transposase and synaptic complex formation. DNA binding is a limiting step for transposition and is subject to tight regulation. Firstly, the transposase has to sample the genome to find the terminal ends of the transposon. There are three general mechanisms of how an enzyme can find its DNA target, such as short diffusive hopping motions, sliding along the chromosome or by direct transfer to the target DNA sequence (Halford and Marko 2004). It has been suggested that Tnp localizes transposon inverted repeats through direct transfer mechanism, although no strong evidence supports this hypothesis (Steiniger *et al.*, 2006).

It has often been proposed that Tn*5* has evolved sub-optimal properties to bind DNA as an auto-regulatory strategy. Tnp has low affinity for its OE or IE sequences. In fact, an artificial sequence that merge both terminal ends, named *mosaic ends* (ME), has been developed which increase transposition rate *in vivo* (Zhou *et al.*, 1998). Two pre-cleaved ME sequences have been used together with an hyperactive Tnp to form an experimental analog of the synaptic complex (termed paired-end complex), and the X-ray co-crystal structure have been resolved (Davies *et al.*, 2000).

The DNA binding domain is located at the N-terminus of Tn*5* (Weinreich *et al.*, 1994b), while the C-terminus of Tnp contains the dimerization domain necessary for transpososome formation (Steiniger-White and Reznikoff 2000). In solution Tnp is found mainly as a monomer but it only binds DNA as a dimer (Braam *et al.*, 1999). Functional and structural studies lead to a model where the C and N-terminus are in close contact limiting the ability of Tnp to bind DNA nor form the synaptic complex (Reznikoff 2008). It has been hypothesized that the N-terminal binds to DNA during translation, when it is being synthesized and the N domain is still free. This could contribute to the *in vivo* bias observed in *cis* in Wt Tnp transposition (Reznikoff 2008). Unfortunately, there is no biochemical evidence to support this hypothesis in Tn*5*.

Additionally to its auto-regulated mechanism, Tnp is down-regulated by the inhibitory protein. Inh is encoded in the same reading frame than Tnp but transcribed form its own promoters and translated from a distinct initiation codon. Its structure is known in atomic detail (Davies *et al.*, 1999). Inh is unable to bind DNA by itself, but heterodimerizes with Tnp, promoting unspecific DNA binding (De la Cruz *et al.*, 1993; Braam *et al.*, 1999) and prevents Tnp to find the OE.

1.7. With a little help from the host

Transposition could also be regulated by host-imposed mechanisms. For example, general host processes like Dam methylation could influence in transcriptional control in ISs. It has been reported that hemi-methylated ends of Tn*10* or Tn*5* are more active than fully methylated forms (Roberts *et al.*, 1985; Yin *et al.*, 1988). This implies that Tn*10* and Tn*5* are only activated with the passage of the replication fork when the chromosome is transitorily hemimethylated.

On the other hand, although highly autonomous, ISs may require host factors to facilitate or make possible transposition. Since ISs usually generates nicks or gaps when inserting within target DNA, they require host repair enzymes and other factors (Nagy and

Chandler 2004). It has been proposed that activities of polymerase I are necessary for Tn*5* transposition *in vivo* (Sasakawa *et al.*, 1981). Other enzymes that modify the structure or topology of the DNA such as gyrase (Isberg and Syvanen 1982) or topoisomerase I (Sternglanz *et al.*, 1981) could influence in transposition. All these examples point out to a possible association between replication and transposition.

Nevertheless, little evidence has been reported of direct interaction between a transposase and a host factor. A study on Tn*7* transposon showed that TnsE, a Tn*7*-encoded factor that targets transposition preferentially to replicating conjugative plasmids, interacts with the β sliding clamp (Parks *et al.*, 2009). β is an essential replication factor that provides processivity to DNA polymerases and coordinates numerous enzymatic activities in the replisome (López de Saro *et al.*, 2003; Johnson and O'Donnell 2005). However, Tn*7* encodes five proteins, and it was unclear whether interactions with the replisome could be generalized to transposases of less complex mobile elements.

Because the known transposition pathways often require host enzymatic functions including DNA polymerases and other factors implicated in DNA replication, it is possible that transposition takes place concurrently with chromosomal replication. In this work we investigate what biochemical mechanisms could be behind the association between replication and transposition, or in other words how a transposase can target the replication fork. Although some discrete examples have been presented (Parks *et al.*, 2009), no general mechanism linking these processes has been proposed.

1.8. The replication fork: structure and organization

The DNA replication is a highly complex process that requires the assembly and coordination of a large number of enzymatic activities to ensure the faithful copy of the genome itself. The initiation of replication is a tightly controlled mechanism that involves the recognition of a discrete region in the circular bacterial chromosomes named oriC by a complex of proteins, promoting the transiently separation of parental strands (Mott and Berger 2007). Replication elongation is undertaken by another multi-protein complex, termed the replisome, that associated with a particular structure of the DNA are collectively known as the replication fork. Elongation proceeds bidirectional and simultaneously in both halves of the chromosome (i.e. replichores) by the advance of the replication fork (Kornberg and Baker 1992). As the replication fork moves along the DNA, parental strands are unwound and daughter strands are synthesized. The new DNA strand is synthesized in the 5´to 3´direction. Because of the antiparallel nature of DNA, one

strand is synthesized continuously (leading strand) while in the other strand (lagging strand) the 5´to 3´ DNA synthesis proceeds discontinuously as a series of Okazaki fragments, in the relative direction opposite to the advance of the replication fork. In *E. coli* replication ends at *ter* sites where terminator protein binds, arresting the movement of the replication fork. Finally, parental and newly synthesized chromosomes are resolved by site-specific recombinases (Blakely *et al.*, 1993).

The replisome of *E. coli* is formed by a diverse collection of enzymes which activities are precisely synchronized. Parental DNA is first unwound by DNA helicase, and the resulting topological stress is relieved by topoisomerases. Short RNA fragments synthesized by a primase are used as a primer by the core subunit of the holo-enzyme DNA polymerase III (Pol III) to create the new DNA strand (Davey and O'Donnell 2000). The processivity of the polymerase (i.e. the tendency of the enzyme to remain on their template rather than dissociate and associate) is enhanced by its association with β sliding clamp that encircles the DNA and tether polymerase to the template during replication (Fig. 1.5; Georgescu *et al.*, 2008).

The interplay between chromosomal replication and other processes, such as transcription or chromosomal segregation, shapes the organization of bacterial chromosomes (Rocha 2008; Sobetzko *et al.*, 2012). It can result in specific patterns of localization or orientation of genes in the chromosome relative to the origin of replication and the direction of advance of the replication fork. For example, highly expressed genes tend to cluster near the origin of replication in fast replicating bacteria (Couturier and Rocha 2006), and essential operons like those encoding the highly expressed *rrn* genes tend to be placed in the leading strand, possibly to prevent the instability caused by head-on clashes between the replication and transcription machineries (Rocha and Danchin 2003; Srivatsan *et al.*, 2010; Paul *et al.*, 2013).

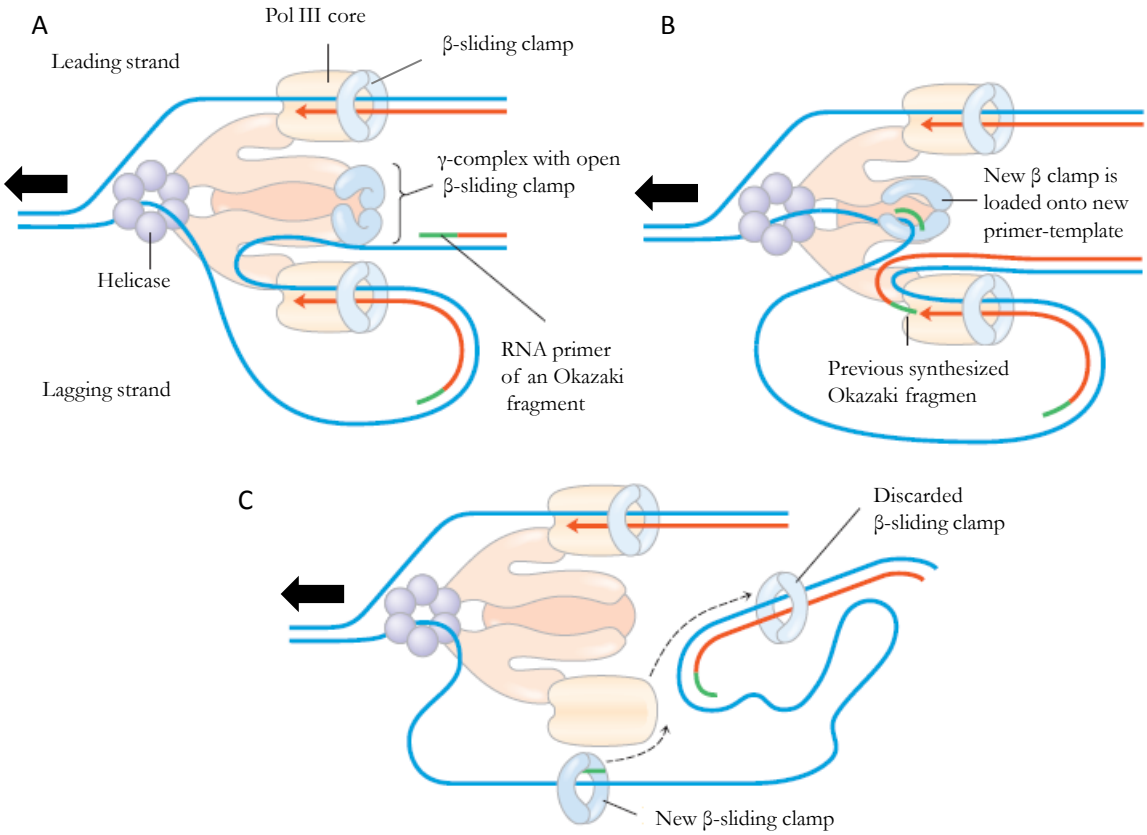


Figure 1.5. Schematic representation of the replication fork.
A) Parental DNA strands (blue lines) are unwound by a helicase as the replication fork advance (black arrows indicate the movement of the fork). Replication is coordinated by the holo-enzyme DNA polymerase III, where the core enzyme (Pol III core) is the responsible of the new DNA synthesis. Red arrows indicate the 3´end and the direction of synthesis of the new strands of DNA. The leading strand is synthesized continuously and the lagging strand discontinuously, as a series of Okazaki fragment (in green RNA primer of each Okazaki fragment). B) The β sliding clamp is loaded and unloaded onto DNA (dotted lines in C) by the γ-complex. C) In the lagging strand after concluding the synthesis of a Okazaki fragment, β clamp remain transitory bound to DNA. Reproduced from Nelson and Cox 2004.

1.9. Sliding clamps

1.9.1 Sliding clamps are conserved and universal replication factors

Sliding clamps are ancient proteins with a key role in the development and evolution of DNA biochemical mechanisms involved in genome replication and stability. In fact it has been proposed that sliding clamps were even present in the last universal common ancestor (LUCA), while others enzymes like helicase, primase or DNA polymerase were not (Georgescu *et al.*, 2015). Thus, sliding clamps are highly conserved in the three branches of the tree of life, although the amino acid sequences between prokaryotic and eukaryotic clamps show no homology (Indiani and O'Donnell 2006). However, all of them share a ring-shaped structure, formed by homo-oligomers that encircle and interact with DNA in a

topological manner, which allow clamps to freely slide along the chromosome, serving as a binding platform for other proteins, including DNA polymerases (Bruck and O'Donnell 2001).

Crystallographic studies revealed that the β sliding clamp in bacteria adopts a homo-dimeric structure in solution. Each 40 kDa monomer is constructed from three globular domains with the same folding conformation (Kong *et al.*, 1992). β monomers are assembled head-to-tail and as a result of this arrangement β dimer has two different sides, the amino- and the carboxyl-faces (Georgescu *et al.*, 2008). On the other hand, the archaeal and eukaryotic counterpart, called proliferating cell nuclear antigen (PCNA), is a homo-trimer. The three PCNA monomers, consisting in two globular domains each of them, are also aligned head-to-tail to form a ring-shaped protein of the size of β dimer (Fig. 1.6; Krishna *et al.*, 1994; Matsumiya *et al.*, 2001). Interestingly, sliding clamp-like proteins have also evolved in some viruses. The gen 45 protein (gp 45) of bacteriophage T4 is a homo-trimer protein that encompasses DNA, giving processivity to the viral DNA polymerase gp 43 (Moarefi *et al.*, 2000).

Since all sliding clamps are closed circular structures, they cannot assemble by themselves on DNA. Instead, they require a multiple subunit protein complex known as clamp loader to open the ring and close it around the DNA in an ATP-dependent enzymatic reaction (Jeruzalmi *et al.*, 2001a). Clamp loaders not only assembled clamps on DNA, but also they target them to sites where DNA is been synthesized, loading clamps around a 3´ end of a primer-template junctions. Moreover, clamps are loaded on DNA in a defined orientation to interact with DNA polymerases and other enzymes (Hedglin *et al.*, 2013). The processivity clamp loader γ-complex of *E. coli,* consists in five different kind of subunits with a stoichiometry of $\gamma_3\delta\delta'\chi\psi$ (Jeruzalmi *et al.*, 2001b; Pritchard *et al.*, 2000) where $\gamma_3\delta\delta'$ are sufficient to load β onto DNA (Onrust *et al.*, 1991). Clamp loader architectures are also evolutionary conserved as revealed by structural studies of the eukaryotic and archaeal replicative factor C (RFC) (Bowman *et al.*, 2004; Pisani *et al.*, 2000) and the gp44/62 clamp loader complex of the bacteriophage T4 (Jarvis *et al.*, 1989).

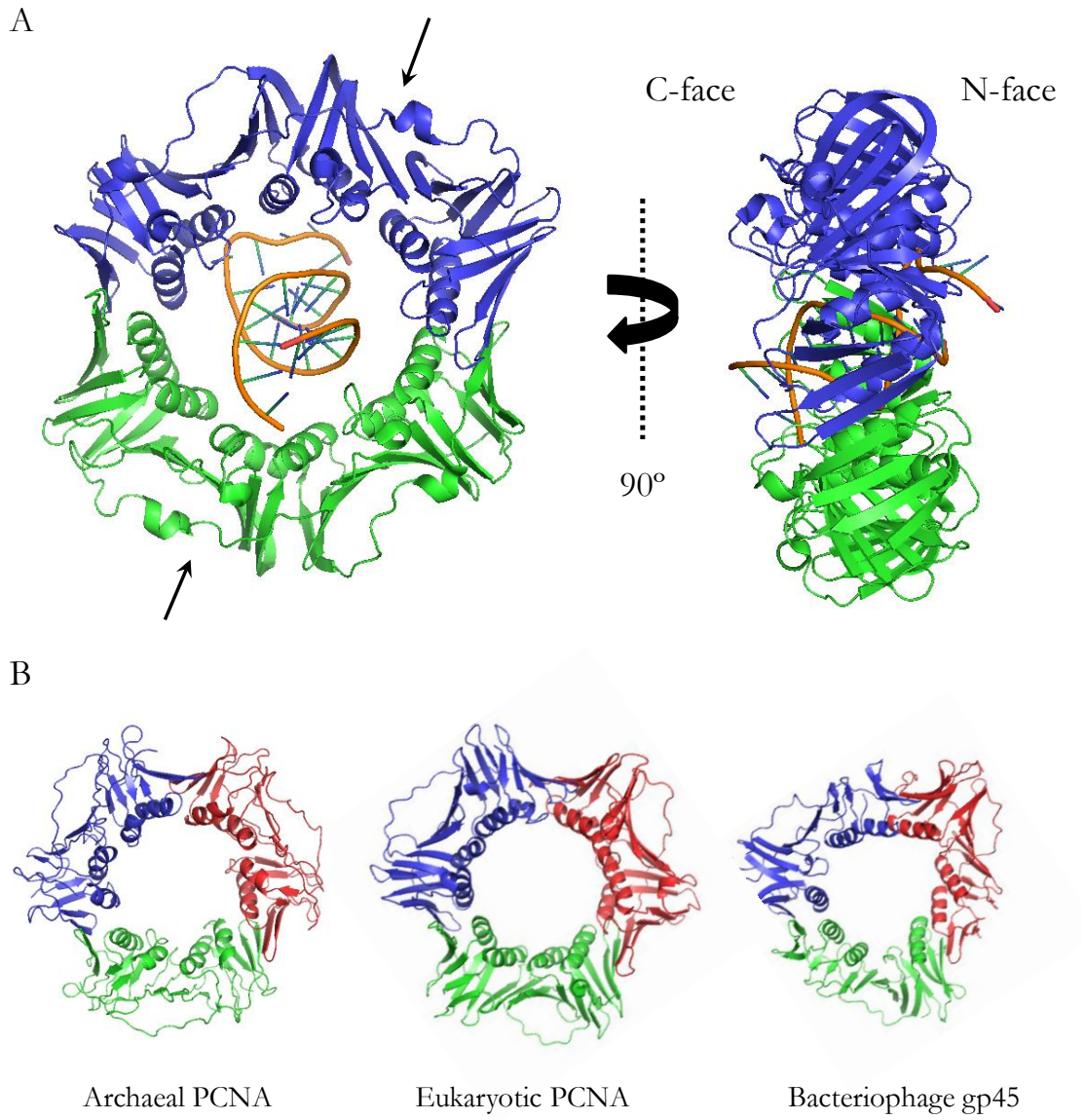


Figure 1.6. The structure of sliding clamps is evolutionary conserved.
A) Crystallographic structure of *Escherichia coli* β sliding clamp. Front (left) and side (right) views are represented (PDB ID: 2POL). Arrows indicate the hydrophobic pockets on the surface of each monomer that are the sites of interaction of all β partners studied. DNA is represented in the cavity formed by the two β monomers. B) Comparative structures of representative sliding clamps. From left to right, archaeal PCNA (3HI8), eukaryotic PCNA (1SXJ) and gp45 of a bacteriophage (1CZD). Monomers of each sliding clamp are represented in different colors.

1.9.2. Sliding clamps coordinate diverse DNA biochemical mechanisms

The assembly of the ring-shaped conformation of sliding clamps defines two distinct faces in their structure. The interaction of clamps with their partners takes place at hydrophobic pocket located in the carboxyl-face of each monomer and therefore the β dimer has two binding sites and PCNA trimer has three (Hedglin *et al.*, 2013). Clamps are co-factors of diverse proteins involved in DNA metabolism mechanisms. Most sliding clamp ligands

analyzed so far interact with clamps via a binding residues located preferentially in the C-terminal regions of the protein, although N-terminal or internal binding sites could also be found (López de Saro 2009). A survey of the binding sequences in β-interacting proteins revealed that all share a short and poorly conserved pentapeptide motif enriched in hydrophobic amino acids which often resides in unstructured regions of the protein. The consensus sequence of the binding motif is Q-L-S/D-L-F, being the most strongly conserved residues Q_1 and L_4, although it accepts some degree of looseness (Dalrymple *et al.*, 2001). On the other hand, PCNA partners share a consensus interacting motif (Q-x-x-L-x-x-F-F) which sequence resemble to some extent the canonical β binding motif (Moldovan *et al.*, 2007).

β and PCNA are quite stable on DNA, requiring clamp loaders to recycle them on and off DNA during replication (Yao *et al.*, 1996). Indeed, recent *in vivo* stoichiometry studies of the replisome of *E. coli* and *B. subtilis*, have revealed that β is slowly recycled after synthesis of the Okazaki fragments and tends to accumulate in highly condensed "clamp zones" in the lagging strand (Su'etsugu and Errington 2011; Moolman *et al.*, 2014). β in "clamp zones" are presumably free to interact with other factors involved in DNA biochemistry, serving as a platform for the recruitment of enzymes after the passage of the replication fork. In fact, it has been proposed that clamps originally evolved to mark newly synthesized DNA and recruit enzymes required for genome stability, rather than been a processivity factor (Georgescu *et al.*, 2015). Hence, clamps target enzymes involved in DNA metabolism, like maturation of Okazaki fragments, DNA repair or cell cycle control (Warbrick *et al.*, 2000). For instance, enzymes of the mismatch repair system in bacteria, MutL and MutS, both interact with β (López de Saro and O'Donnell 2001; López de Saro *et al.*, 2006), and their eukaryotic homologues, MSH3 and MSH6, also bind to PCNA (Warbrick *et al.*, 2000). The interaction with sliding clamps could be a critical step of the activity regulation of those enzymes which clamps are essential co-factors. Since most of the described β-interacting proteins bind competitively in the same hydrophobic pocket of β, the relative binding affinity of the sliding clamp for its partners could determine how these processes are orchestrated and regulated (López de Saro *et al.*, 2003). Moreover, many proteins that harbor binding motifs belong to unrelated structural families, which reflect convergent evolution for interacting with sliding clamps as a way to target the replication fork (López de Saro 2009).

Objectives

2. Objectives

This thesis aims to gain insight into the interplay between transposition and host replication functions. Particular objectives pursued are:

1. To survey specific patterns of orientation of IS genes in chromosomes related to the replication machinery.
2. To identify transposases of diverse IS families that interact with the replication factor β sliding clamp.
3. To characterize and map the interaction of sliding clamps with purified transposases.
4. To develop tools that allow the study of dynamics and evolution of ISs population within a genome and in environmental samples.
5. To investigate the effects of transposase interaction with β sliding clamp on IS proliferation and transposition rate.

Materials and Methods

Unless otherwise indicated, described methods were performed at room temperature and procedures that involved the use of commercial supplies followed manufacturer's recommendations. Regular procedures of molecular cloning techniques and protein manipulation were carried out accordingly to standard protocols (Sambrook *et al.* 2001).

3.1. Oligonucleotides and peptides

Oligonucleotide sequences used for gene amplification, cloning and mutagenesis are listed in Table 3.1

ID	Oligonucleotide sequence (5´to 3´)	Use
1	CGGCTAGCCATGGAATTTACCGTAGAACGTGAGC	*Eco* β amplification
2	GGCCTAGGATCCTTACAGTCTCATTGGCATGACAACATAAGCCGC	〃
3	CTGGTCCCATGGCGCTCAAGGCGGATCGGGCAACG	*Apm* β amplification
4	CTGGTCGGATCCTCAGACCCGCATCGGCATGAGAACGAAAATCG	〃
5	CGGCTAGCTAGCGACTACAAGGACGACGATGACAAAGAAATTATCGTTGATCAGGAGAC	*Lfe* β amplification
6	GGCCTAAGATCTTCACCCCATGGGCATTAACACATACCGGG	〃
7	GGATGACCATGGTTAAGGCAGCAATTAATGCAGAGCTTC	*Mba* PCNA amplification
8	GGATGAGGATCCTCAGTCCGACTCAATTCTTGGAGCCAGG	〃
9	GGACATCGTACATGTTACGCCGCGCAAGCGTTGGTTTCGAGGCGCGCCTGGTC	*h*PCNA amplification
10	GCTCAGCAGGGATCCTTATTACTAAGATCCTTCTTCATCCTCG	〃
11	GCGGTGCCCATGGGCAGTTATCAGGTCTTAGCCCG	*E. coli* γ amplification
12	GCGGTGCGGATCCTCAATGATGATGATGATGTGGCTCAGGCAGCGGCATACGCGGATGG	〃
13	GGCGACGACATATGATTCGGTTGTACCCGGAACAACTCC	*E. coli* δ amplification
14	CGTCGCCGGATCCTCAACCGTCGATAAATACGTCCGCCAGGGG	〃
15	CGGCGACGCATATGAGATGGTATCCATGGTTACGACC	*E. coli* δ´ amplification
16	CGTCGTCGGATCCTTAAAGATGAGGAACCGGTAGCACAACGCCCGG	〃
17	GAATGACGGATCCGTCGGCGTGGAACGCACGATGGCGGAAG	PolIV^{LF} amplification
18	GAATGACCTCGAGTCATAATCCCAGCACCAGTTGTCTTTCCATTTGCGGG	〃
19	GCCGACGGATCCCTAATAAAGAAAAAAAGAG	LINE-1^{Z} amplification
20	GCGATCCTCGAGTAGAATTCGGCTGTGAATCC	〃
21	CACCGATCCCACAGAAATAGCGACTACCATCAGAGAATAC	LINE-1^{PIP1} mutagenesis
22	GTATTCTCTGATGGTAGTCGCTATTTCTGTGGGATCGGTG	〃
23	CAAACTACCATCAGAGAAGCCGCCAAACACCTCTACGCAAATAAAC	LINE-1^{PIP2} mutagenesis
24	GTTTATTTGCGTAGAGGTGTTTGGCGGCTTCTCTGATGGTAGTTTG	〃
25	GCGAACCATGGTCATCGACGTGGTTCCGAATGGCC	IS*1634*Tnp amplification
26	GCCATACCCATATGCTACTGGACACGCATGGGTTCG	〃
27	TCGAGGAGGTTGGCGGCGGCCGCCGCTAGTGTGGTTGG	IS*1634* 5Δ mutagenesis
28	CCAACCACACTAGCGGCGGCCGCCGCCAACCTCCTCGA	〃
29	ATGGGTTCGATTTCGAGGAGGTTGAATAGGCTTAGCTGTAGTGTGGTTGGCGTGGCCAGGG	IS*1634* cn mutagenesis
30	CCCTGGCCACGCCAACCACACTACAGCTAAGCCTATTCAACCTCCTCGAAATCGAACCCAT	〃
31	CGGCTGCCATGGTAACTTCTGCTCTTCATCGTGCG	Tn*5* Tnp amplification
32	GGCCTAGAATTCTCAGATCTTGATCCCCTGCGCCATCAG	〃
33	GGCTCATCCATGGCGGCCGACTGGGCTAAATCTGTG	Tnp$^{N\Delta7}$ amplification
34	GCACCGGGATCCTCAGATCTTGATCCCCTGCGC	〃

35	GGCTCATCCATGGCGGAAGGCGCTTACCGATTTATCCGC	Inh amplification
36	GGCCTAGAATTCTCAGATCTTGATCCCCTGCGCCATCAG	〃
37	CTTTCTCTGAGCTGTAACGCCCTGACCGCAACAAACGA	Tnp^{L363A} mutagenesis
38	TCGTTTGTTGCGGTCAGGGCGTTACAGCTCAGAGAAAG	〃
39	CGTGAAGCTTTCTCTGAACTGTAACAGCCTGACCG	Tnp^{L366F} mutagenesis
40	CGGTCAGGCTGTTACAGTTCAGAGAAAGCTTCACG	〃
41	GCCATCCAGTTTACTTTACAGGGCTTCCCAACCTT	$Tnp^{C\Delta 20}$ mutagenesis
42	AAGGTTGGGAAGCCCTGTAAAGTAAACTGGATGGC	〃
43	GCGAACCATGGTCATCGACGTGGTTCCGAATGGCC	pSKT1- IS*1634* cloning
44	GCTACCGAATTCCTACTGGACACGCATGGGTTCG	〃
45	GGCCGTTAAAACTTCGGACTAGTACTCGAG	pSKT1-IS*1634* RIR cloning
46	CTAGCTCGAGTACTAGTCCGAAGTTTTAAC	〃
47	CTAGCTCGAGGACTAGTCCGAAGTTTTTAT	pSKT1-IS*1634* LIR cloning
48	CCGGATAAAAACTTCGGACTAGTCCTCGAG	〃
49	CGGCTGCCATGGTAACTTCTGCTCTTCATCGTGCG	pSKT1-Tn*5* cloning
50	GGCCTAGAATTCTCAGATCTTGATCCCCTGCGCCATCAG	〃
51	CTAGCTCGAGCTGTCTCTTGTACACATCT	pSKT1-Tn*5* RIR cloning
52	CCGGAGATGTGTACAAGAGACAGCTCGAG	〃
53	CTAGCTCGAGCTGTCTCTTGTACACATCT	pSKT1-Tn*5* LIR cloning
54	GGCCAGATGTGTACAAGAGACAGCTCGAG	〃
55	AAGGCACGATCGCCAATCTGA	IS*1634* qPCR
56	CAGCAGGCGCGGAATCTCCAG	〃
57	CGTCCTCGCCCGCAAATACC	*Apm* dnaX qPCR
58	AATCCGCGCCGTCGTCGTCTT	〃
59	GCTCTCAAGCTTTGCCGACCTGATGGCGCATC	IS*1634 i*PCR
60	GCCATCGACCAGCTCCGCCGGCATCTGGCTCAG	〃

Table 3.1. List of oligonucleotide sequences

Oligonucleotide sequences of *Acidiphilium sp PM* transposase and related genes used in the microarray are listed in Table III.1 in Appendix III. Peptide sequences used in biochemical assay are shown in Table 3.2

Name	Peptide sequence (N-ter to C-ter)	Organism
Pol IV (NP_414766)	VGLHVTLLDPQMERQLVLGL	*Escherichia coli*
Pol IV m	VGLHVTLLDPQMERALVAGL	〃
IS*5*a (AAB53644)	QIQGVAENDNQLAMLFTLAN	〃
IS*5*a m	QIQGVAENDNALAMAFTLAN	〃
IS*30* (NP_415922)	YFPKKTCLAQYTQHELDLVA	〃
IS*30* m	YFPKKTCLAQYTAHEADLVA	〃
IS*66* TnpB (YP_424826)	RDGKVHLTPAQLSMLLEGIN	〃

IS*66* TnpB m	RDGKVHLTPAALSMALEGIN	〃
IS*66* TnpC a (YP_003235004)	SEQAEALRQKDQQLSLVEET	〃
IS*66* TnpCa m	SEQAEALRQKDQALSAVEET	〃
IS*66* TnpC b (ZP_07592975)	RFGKKCESLAGMQRSLFEED	〃
IS*66* TnpCb m	RFGKKCESLAGMARSAFEED	〃
IS*91* (ACO24927)	ERAPPLTPSLFDPSQSRLFD	〃
IS*91* m	ERAPPLTPSLFDPSASRAFD	〃
IS*200* (1) (ZP_03029803)	YARYQEKMEQTHEQQMELLE	〃
IS*200* (1) m	YARYQEKMEQTHEQAMEALE	〃
IS*200* (2) (NP_752024)	YIKHQLEEDKMGEQLSIPYP	〃
IS*200* (2) m	YIKHQLEEDKMGEALSAPYP	〃
IS*1380* (YP_003829282)	VLKPEKERAQLSLLEGSEYD	〃
IS*1380* m	VLKPEKERAALSALEGSEYD	〃
ISL*3* (ZP_07122173)	MCEKEPELKIAQQLVLEFYR	〃
ISL*3* m	MCEKEPELKIAQALVAEFYR	〃
ISNCYa (EIL58166)	DVAEMANLPLAEIDKVINLI	〃
ISNCYa m	DVAEMANLPLAEIDAVINAI	〃
Tn*7* TnsC (EIL57895)	GPESEAYDRFKQAGLILDLR	〃
Tn7 TnsC m	GPESEAYDRFKAAGAILDLR	〃
MbaIS*200* (YP_307176)	NQGNQEEKEAYKQMKIIDFQ	*Methanosarcina barkeri*
MbaIS*200* m	NQGNQEEKEAYKAMKIIDAQ	〃
MbaIS*1634*(WP_011306730)	DEIKSKIIRLMGK	〃
MbaIS*1634* m	DEIKSKIIRAAGK	〃
ApmIS*1634 (*WP_007423974)	ATPTTLQAAAFNLLE	*Acidiphilium sp. PM*
ApmIS*1634* 5a	ATPTTLAAAAANLLE	〃
ApmIS*1634* CN	ATPTTLQLSLFNLLE	〃
Af IS*1634* (WP_014029679)	YDDPAASQQLTLL	*Acidithiobacillus ferrivorans*
Af IS*1634* m	YDDPAASQALTAL	〃

Table 3.2. List of peptide sequences used in biochemical assays
Peptides are N-biotinylated. The NCBI accession number of the protein from which the peptide is derived is provided.

3.2. Microbiological techniques

Escherichia coli strains used in this work were *DH5a* (Thermo Fisher Scientific) for routine maintenance of plasmids, XL10-Gold (Agilent Technologies) for site-directed mutagenesis transformation and BL21 (DE3) (Merck Millipore) or Rossetta2 (DE3) pLysS (Merck Millipore) for protein overexpression. Cells were cultivated in Luria-Bertani (LB) medium (Pronadisa) at 30ºC or in solid LB-agar medium (Pronadisa) at 37ºC. When required, mediums were supplemented with selective antibiotics at concentration of ampilicillin 10 µg/ml, chloramphenicol 35 µg/ml or kanamycin 50 µg/ml.

Competent cells for chemical transformation were prepared by calcium chloride treatment, meanwhile, cells for transformation by electroporation were deionized water treated (Sambrook *et al.* 2001). Briefly, plasmids were incubated along with competent cells on ice for 20 min. Subsequently, in the chemical transformation procedure, cells were subjected to 42ºC for 40 s and afterwards chilled again for 2 min. Alternatively, a pulse of 2000V for 5 ms was applied over electrocompetent cells in a Electroporator 2510 (Eppendorf). Immediately after, in both transformation protocols, cells were resuspended in SOC medium (Invitrogen), outgrow for 1 h at 37ºC and plated on selective plates (Sambrook *et al.* 2000).

Additionally, *Acidiphilium sp. PM* (DSM 24941) was used to establish a 4 years long-term culture. *Acidiphilium sp. PM* was originally isolated from the Tinto river in Huelva (Spain) (Malki *et al.* 2008) and its genome sequenced (San Martín Úriz *et al.* 2011). The sequenced strain was used to initiate a 100 ml culture in GYE medium, which is composed of a mineral salt solution (0.2% $(NH_4)_2SO_4$, 0.01% KCl, 0.033% $K_2HPO_4 \cdot 3H_2O$, 0.025% $MgSO_4 \cdot 7H_2O$, 0.0014% $Ca(NO_3)_2 \cdot 4H_2O$) supplemented with 0.2% (w/v) glucose and 0.01% (w/v) yeast extract. The pH was adjusted to 2.5 with 1 N H_2SO_4 prior to autoclaving (111 °C, 0.5 atm, 30 min). Cultivation took place at 30 °C with vigorous shaking. Serial transfers (1:50 dilution) were performed in periods of 14 days. Under these conditions the cultures reached stationary phase in 5 days and the culture experienced 9-day-periods in nutrient-deprived medium. The 1:50 dilution allowed ~5.6 generations ($\log_2 50$) per serial dilution, or 584 generations in 4 years. Samples from populations were taken periodically and stored at -80 °C.

3.3 Recombinant DNA techniques

DNA and RNA conventional electrophoresis were, otherwise specified, performed in 1% (w/v) agarose (medium EEO, Pronadisa) gels, run at 80 mV, with TAE as running buffer (40 mM Tris-acetate, 1 mM EDTA). Samples were loaded with 1x DNA loading buffer (Invitrogen). 1Kb plus DNA ladder (Invitrogen) was used as molecular weight standard. Gels were stained with SYBR Green I (Lonza) and visualized on a UV trans-iluminator.

3.3.1 Genomic and plasmids DNA extraction

Genomic DNA from pure bacterial cultures was isolated by using Gnome DNA Isolation Kit (MP Biomedicals). For plasmids extraction, the bacterial clone harboring the plasmid of interest was grown overnight in 5 ml of LB media containing selective antibiotics. Then, the culture was pelleted at 4000 g for 10 min and the plasmid isolated using the QIAprep Spin Miniprep Kit (Qiagen). Concentration and quality of extracted DNA was measured with a ND-1000 Spectrophotometer (NanoDrop Technologies Inc.). When necessary, integrity of DNA was checked out in 1% agarose gel electrophoresis.

3.3.2 RNA isolation

For large RNA preparation from *Acidiphilium sp PM*. 50 ml of pure culture was grown to mid-exponential phase, and cells harvested by centrifugation (15 min, 4000 g). We performed a reiterative procedure of acid phenol/chloroform extraction, isopropanol precipitation, followed by another step of TRIzol/chloroform extraction and isopropanol precipitation. Briefly, cells were resuspended in 1.5 ml of LETS buffer (0.1 M LiCl, 0.01 M Na_2EDTA, 0.01 M Tris-Cl , 0.2 % SDS, pH 7.4) at 75ºC and lysed with glass beads and 1 ml of acid phenol (pH 4). 1 ml of chloroform was added and after centrifugation (10 min, 6000 g, 4ºC), the aqueous phase was recovered. The acid phenol/chloroform extraction was repeated twice. RNA was precipitated from the aqueous solution with the same volume of isopropanol, and resuspended in DEPC-treated water. To ensure the maximum purity of the preparation, another extraction step was performed with 1 ml of TRIzol (Thermo Fisher Scientific) and 0.2 ml of chloroform. RNA was again precipitated with isopropanol, washed with ethanol 70% and stored in DEPC-treated water at -80ºC for further applications.

For RNA purification from environmental samples of Tinto river, 2 l of water was filtered through 0.22 μm (Millipore) using a 50 ml sterile syringe. Filters were stored in fixing solution, RNA later (Ambion) and scrapped with a sterile scalpel. Cells were

recovered by centrifugation and RNA extracted using FastRNA ProSoil-Direct Kit (MP Biomedicals). Due to the low concentration of environmental RNA, an amplification step was required. It was performed using the MessageAmp aRNA Amplification Kit (Applied Biosystems). This procedure linearly amplified the RNA, generating an anti sense RNA (aRNA) through a method based on the promoter and RNA polymerase of the phage T7.

All RNA preparations were further treated with TURBO DNase (Ambion) to eliminate any possible trace of contaminant DNA. Subsequently, RNA concentrations were measured with a ND-1000 Spectrophotometer (NanoDrop Technologies), and its quality checked by a Bioanalyzer 2100 (Agilent Technologies).

3.3.3. Polymerase chain reaction (PCR)

DNA was amplified by the polymerase chain reaction (PCR) in a GeneAmp PCR System 9700 (Applied Biosystems). A typical reaction was performed in 50 µl final volume and contained either 100 ng of genomic DNA or 10 ng of plasmid as a template, 250 µM of each dNTP (Roche), 0.2 µM of each primer (see Table 3.1), reaction buffer containing Mg^{2+} 1.5 mM and 1 U of the suitable DNA polymerase.

For routine and screening PCR reactions, Paq5000 DNA polymerase (Agilent Technologies) was used. The reaction comprise the following steps: an initial denaturation of 95 ºC for 5 min; followed by 30 cycles of denaturation (95ºC, 30 sec), annealing (58 to 69ºC, 30 sec) and extension (72ºC for 30 sec/amplicon Kb), and a final extension step of 10 min at 72ºC. For high fidelity PCR amplifications Herculase II Fusion DNA polymerase (Agilent Technologies) was used. PCR reaction conditions with this enzyme are essentially the same from above, using also extensions of 72ºC for 30 sec/amplicon Kb.

All PCR products were purified with QIAquick PCR purification Kit (Qiagen) and were checked in a 1% agarose gel electrophoresis to dismiss any unspecific amplification product.

3.3.4 DNA cloning

DNA was digested with restriction enzymes (New England Biolabs) using suitable buffers and reaction conditions recommended by the manufacturer. Digested products were purified with QIAquick PCR purification Kit (Qiagen). Ligation of compatible digested vector and insert, was performed in a 20 µl reaction with T4 DNA ligase (New England Biolabs), in a buffer containing 1mM ATP, at 16ºC for 4h. Molar ratio of insert : vector in the ligation reaction was at least 4:1. Ligation mixture was transformed in chemically-

competent *E. coli DH5a* by a heat-shock method. Transformants colonies were screened for the insert by PCR and plasmid construction verified by sequencing.

3.3.5. Site-directed mutagenesis

The collection of mutant proteins was mainly performed by *in vitro* site-directed mutagenesis of the respective wild-type (Wt) plasmid. Mutagenesis was carried out by the QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent Technologies). Primers were designed according to the requirements of the reaction. The cycling parameters of the mutagenesis were as follow: an initial denaturing step of of 95 ºC for 2 min; 18 cycles of denaturation (95ºC, 20 sec), annealing (60ºC, 10 sec) and extension (68ºC for 30 sec/Kb of plasmid length), and a final extension step of 5 min at 68ºC.

Mutagenesis amplification products were treated with DpnI to digest the parental supercoiled dsDNA. Afterwards, reaction products were transformed in XL10-Gold ultracompetent cells (Agilent Technologies) by heat-shock and plated in selective plates. Mutagenized plasmids were validated by sequencing.

3.4 *In vivo* transposition assays

The pSKT1 plasmid (Fig 3.1) was provided by H. Savilahti (University of Turku, Finland). It was used as described (Pajunen *et al.* 2010) to study the *in vivo* transposition activity of Tn*5* Tnp, IS*1634* Tnp and their respective mutants. The transposase enzyme was cloned under the transcriptional control of the BAD promoter, allowing the modulation of its expression by addition of arabinose. pSKT1 also harbors a truncated version of the lacZ gene, lacking the first 8 amino acids of the β-galactosidase enzyme. This enzyme is flanked by the cloning site of the inverted repeats of each IS of interest. Chloramphenicol resistance gene was used with selective purposes.

pSKT1-IS*1634*wt was constructed by cloning the transposase gene as a NcoI/EcoRI PCR product. It was amplified from *Acidiphilium sp. PM.* genomic DNA using oligos #43 and #44. Right inverted repeat (RIR) was cloned in pSKT1 using oligos #45, #46, and left inverted repeat (LIR) with oligos #47 and #48 (see Table 3.1). pSKT1-IS*1634*-5A and pSKT1-IS*1634*-cn were constructed by site-directed mutagenesis from the wt plasmid.

In the other hand, pSKT1-Tn*5* was constructed by cloning the transposase gene as a NcoI/EcoRI PCR fragment using oligos #49 and #50. RIR and LIR were cloned using

oligos #51, #52 and #53, #54 respectively. Mutant variants L363A and L366F were also designed by site-directed mutagenesis from the wt plasmid.

In both pSKT1-IS*1634* and pSKT1-Tn*5* assays, a vector with no transposase cloned was used as negative control. pSKT1 constructions were transformed by electroporation in *E. coli DH5a*. A single colony of pSKT1-control, Wt and mutant transposase transformants were spotted in the same plate for *in vivo* transposition analysis. Plates contained LB-agar medium (Difco) supplemented with 100 μg/ml ampicillin (Sigma-Aldrich), 20 μg/ml chloramphenicol (Sigma-Aldrich), 0.05% (w/v) lactose (Sigma-Aldrich), 40 μg/ml of 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) and either 0.1% (w/v) arabinose (Sigma-Aldrich) for pSKT1-*IS1634* or 0.01% arabinose for pSKT1-*Tn5*. Plates with pSKT1-IS*1634* constructions were incubated for 15 days at 30ºC meanwhile pSKT1-Tn*5* and its variant were incubated at 37ºC for 4 days. Four technical replicates and two biological replicates were performed for both set of constructions.

In vivo transposition activity was measured by counting the blue papillae number observed in each transformant colony. Each papilla reflected a transposase-mediated transposition event, generated by the mobilization of the truncated lacZ gene flanked by inverted repeats, into an expressed gene in the correct orientation and reading frame in the bacterial chromosome.

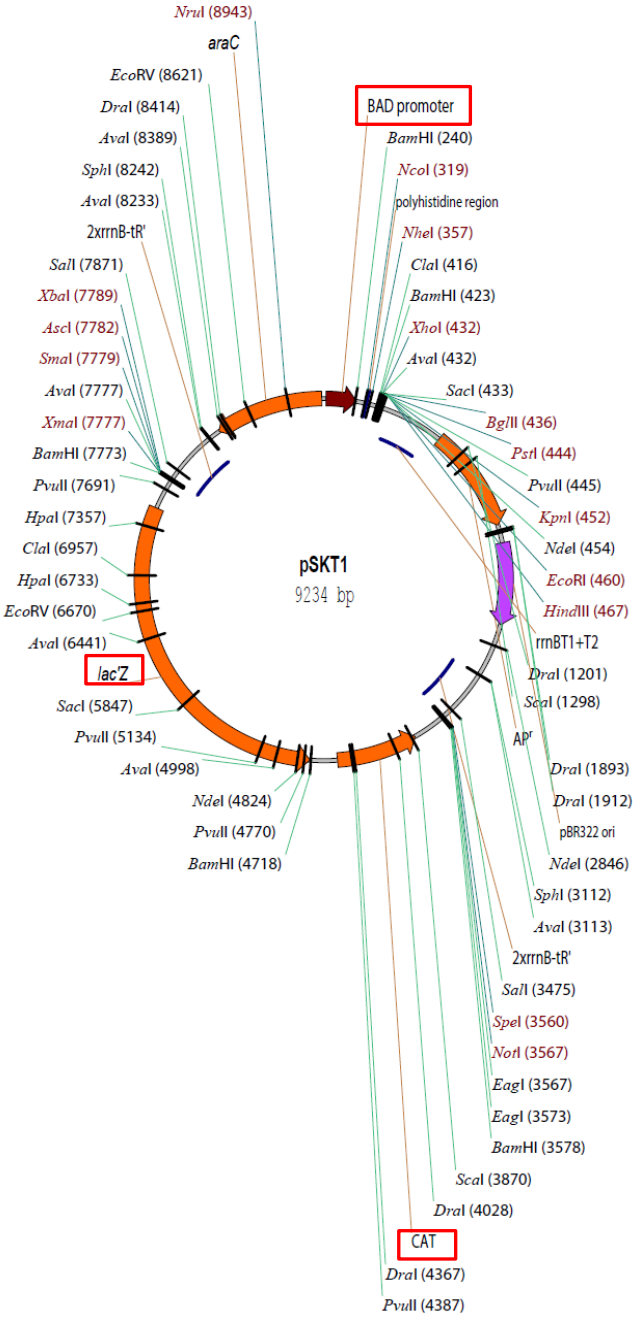


Figure 3.1. pSKT1 plasmid map.
Restriction sites are indicated. The arabinose inducible promoter (BAD promoter), the truncated gene for β-galactosidase (LacZ) and the chloramphenicol resistance gene (CAT), are red boxed. From (Pajunen *et al.* 2010)

3.5 Design and validation of an oligonucleotide-based IS-related genes microarray

3.5.1 Identification of transposases and related genes

To design an oligonucleotide microarray that could detect changes in copy number and gene expression of mobile elements in acidophilic organisms, first we identified transposases and related-genes in 14 fully sequenced acidophilic genomes (see Table 3.3).

Species	Proteins	Transposases & related-elements
Acidiphilium sp.	4252	111
Acidiphilium cryptum JF5	3858	172
Leptospirillum ferrooxidans LfeRT32a	3202	176
L. rubarum	3022	119
L. groupII 5way	3092	90
L. ferrodiazotrophum	3100	142
Acidithiobacillus ferrooxidans 23270	3147	70
A. ferrooxidans 53993	2826	65
A. caldus	2821	134
Acidimicrobium ferrooxidans 10331	1964	62
Ferroplasma acidarmanus fer1	1986	95
Thermoplasma acidophilum	1482	17
Haloquadratum walsbyi	2646	47
Salinibacter ruber	2833	58

Table 3.3. Transposases and related elements detected in selected organisms
First column, organisms which transposases are represented in the microarray (Species). Total hypothetical proteins codified by each organism and transposases identified in those proteins, second and third column respectively.

We used RPS Blast (NCBI) to scan 40231 proteins identified in these genomes with a set of 149 position-specific scoring matrix (PSSM) profiles, selected from Pfam and corresponding to mobile element proteins. This procedure identified a set of 1358 transposases and associated proteins. The corresponding gene sequences were progressively clustered by means of a customized pipeline that used UCLUST (Edgar 2010) and Cons (Rice *et al.* 2000) to generate a set of 769 consensus sequences characterized by having at least two conserved blocks of minimal length equal to 50 bp and zero ambiguities.

3.5.2 Design and construction of the microarray

The program Array Designer 4 (Premier Biosoft International) was used to design two different oligonucleotides for each consensus sequence (1538 oligonucleotides) with an average length of 40 nt and a constant estimated melting temperature (72 °C). See Table III.1 in Appendix III for sequences corresponding to *Acidiphilium* IS-related ORFs. Hence, the construction of the microarray is redundant, because two different regions of the same transposase or related-element are recognized by two different oligos. Importantly, any IS containing more than one ORF (e.g., one that contains the transposase and an accessory element), will be represented by oligonucleotides from each of these ORFs. In addition to the oligonucleotides representing the mobile genes present in the 12 genomes of acidophiles, we included also three reference genes (rpoB, dnaX and gyrB) for each of those genomes. Those genes are in a single copy in the chromosome, are constitutively expressed and are highly conserved. Each reference gene is also represented in the microarray by two different oligonucleotides.

Spotting was carried out with the MicroGrid-TAS II Arrayer (Genomic Solutions) at 22 °C and 50–60% relative humidity on epoxy-substrate slides (Arrayit) according to the manufacturer's instructions. An array containing 10752 spots (including three replicas) was constructed in each slide.

3.5.3 Sample labeling

Genomic DNA obtained from the founder strain ("2007") and from the 4-year-old population ("2011") of the *Acidiphilium* long-term culture, were differentially labeled with cyanine fluorescent dyes, Cy3 and Cy5. First, DNA was sonicated with a digital Sonifier (Branson) until DNA fragments between 2 to 0.8 Kb were reached. Afterwards, 2 µg of genomic DNA was denatured at 95ºC for 10 min. Then, DNA was fluorescently labeled in a reaction containing 50 U of Klenow fragment (New England Biolabs), random hexamer primers (Invitrogen), 0.5 mM dCTP, dATP, dGTP and 0.3 mM of either labeled Cy3-dUTP (Amershan) or Cy5-dUTP (Amershan). Reaction was performed at 37ºC for 2h. Unincorporated nucleotides were removed using QIAquick PCR purification Kit (Qiagen). pmols of incorporated label were calculated using a ND-1000 Spectrophotometer (NanoDrop Technologies).

In the other hand, 25µg of total RNA extracted from pure culture of *Acidiphilium sp PM.,* was used as template for a first strand cDNA synthesis. Reaction was performed in a 20 µl volume, containing 250 µM of each dNTP (Roche), random nonamer primers

(Invitrogen) and 100 U of CyScript reverse transcriptase enzyme (Amershan). Reaction was incubated at 42ºC for 2 h. Resulting cDNA was treated with RNase H (New England Biolabs). Then, 2 µg of cDNA was used as a template for a labeling reaction, using Cy3-dUTP (Amershan) and 50 U of Klenow fragment (New England Biolabs) as previously described. Alternatively, RNA extracted from environmental samples was linearly amplified generating anti-sense RNA (aRNA). aRNA was used as a template in a reverse transcriptase reaction performed with CyScribe First-Strand cDNA Labelling Kit (Amershan) and using Cy3-dUTP (Amershan) as labeled nucleotide.

3.5.4 Microarray hybridization and scanning

First, we aimed to determinate the optimal hybridization conditions and the specificity of our microarray. We labeled genomic DNA of *Acidiphilium sp.* PM with fluorescent dye Cy3-dUTP. Then, we hybridized the labeled DNA against the microarray at two different temperatures, 55 and 65°C. At 65°C we observed optimal specificity because we only detected significant signal for those oligonucleotides representing the *Acidiphilium sp.* reference genes and not for those of other microorganisms (Fig. 3.2)

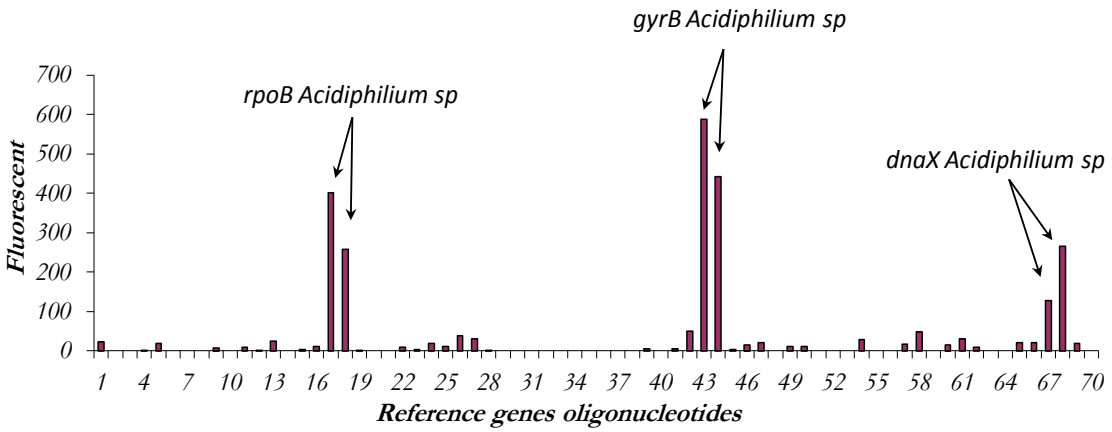


Figure 3.2. Optimal hybridization conditions and specificity of the microarray.
Reference gene probes that are detected when *Acidiphilium sp.* PM genomic DNA is hybridized against the microarray at 65ºC. *Acidiphilium* reference genes are indicated with arrows. Fluorescent is given in arbitrary units.

Hybridization procedure of the microarray was performed as follow. The microarray was denatured 2 min at 94ºC, washed in distilled water and fixed in cold absolute ethanol. Subsequently, it was blocked at 42ºC for 1 h in pre-hybridation buffer (5x SSC, 0.1% SDS,

0.1% BSA, 0.1 mg ml^{-1} denatured herring sperm DNA). Equimolar amounts of differentially labeled genomic DNA from "2007" and "2011" *Acidiphilium sp.* strains, with at least 25 pmols of incorporated label, were hybridized competitively against the microarray. DNA was denatured at 94ºC for 2 min and hybridized in the microarray in HybIt hybridization solution (Arrayit) in a hybridization chamber (Arrayit) at 65°C overnight. Labeled cDNA samples were hybridized as above.

After hybridization, slides were profusely washed three times for 5 min, first in buffer containing 2x SSC and 0.1% SDS; then in less stringent buffer (0.2x SSC, 0.1% SDS); and finally in 0.2x SSC buffer. The slide was dried by centrifugation on a Galaxy Miniarray slide centrifuge (VWR).

Microarray was scanned in a GenePix 4100A scanner (Axon Instruments) at 800 gain, using excitation wavelength of 532 nm and 635 nm. Resulting images were analyzed with Genepix pro v.6.0 software (Axon Instruments) by measuring the fluorescence intensity in each spot. Local median background was subtracted from the median intensity in each spot. Results were normalized by the median of ratios of the reference genes. The transposase copy number change was represented as the $\log_2$ of the ratio intensities [$\log_2$("2011"/"2007")]. $\log_2 > 0.5$ implied increase in transposase copy number at the end of the long-term culture. Conversely negative $\log_2$ indicated loss of transposase genes.

Results of the microarray represented in Figure 4.16 C and Table III.1 are the average change of two different oligonucleotides representing a single transposase and replicated both of them three times in each microarray. The microarray were assayed three independent times, one of them switching the fluorescent dyes Cy3 and Cy5 used to label the DNA from the two strains. For expression microarrays, fluorescence arbitrary units of the transcription level detected for each transposase, was represented.

3.5.5 Microarray validation

Since microarrays are semi-quantitative techniques, we used two different approaches to confirm the increment in genomic copy number of IS*1634* detected with the microarray at the end of the long-term culture.

3.5.5.1 Quantitative PCR

DNA was extracted from both the initial and final time-points of the long-term *Acidiphilium sp.* PM pure culture. The concentration and quality of extracted DNA was measured by a Qubit Fluorometer (Life Technologies). Primer sequences used in the qPCR were oligo

#55 and oligo #56 for the *Acidiphilium* IS*1634* transposase gene; and oligo #57 and #58 for the dnaX gene (See Table 3.1). PCR reactions (25 µl) were performed with 0.4 µM primers (each), *Acidiphilium sp. PM* "2011" target genomic DNA at two different final concentrations (40 pg/µl and 8 pg/µl) and 12.5 µl of iQ SYBR Green Supermix (Life Technologies) according to the manufacturer's instructions. Two biological and three technical replicates were prepared for each gene and each DNA concentration. Reactions were carried out with a MyiQ Single-Colour Real-Time PCR Detection System (Bio-Rad). Cycling parameters comprised an initial cycle of 3 minutes at 95 °C followed by 40 cycles of 30 sec at 95 °C, 30 sec at 60 °C and 30 sec at 72 °C. A melting curve analysis was performed for each reaction to rule out non specific reaction products or primer dimers. "2007" genomic DNA was serially diluted in 5-fold increments from 5 ng to 1, 6 pg. These serial dilutions with three technical replicates, were used to create a standard curve for the reference dnaX gene (r = 0.992) and another curve for IS*1634* transposase gene (r = 0.997). The genomic copy numbers of *Acidiphilium* IS*1634* and dnaX in "2011" time-point was determined from their corresponding "2007" standard curves where both genes are present in a single copy per chromosome, using the comparative threshold cycle method (CT). Results in Fig. 4.17 A represent the average of "2011" dnaX and *Acidiphilium* IS*1634* chromosomal gene copy number and the standard deviation (SD) of the replicates.

3.5.5.2 Inverse PCR

Inverse PCR was used to amplify the flanking regions of the *Acidiphilium* IS*1634* insertion sites detected in the "2011" culture. For this purpose 1 µg of "2007" and "2011" *Acidiphilium sp.* PM genomic DNA were digested with EcoRI (NEB), an enzyme that does not cut within the transposase sequence, for 2 h at 37 °C in 20 µl. Digestions were ligated in 200 µl reaction volume with T4 ligase (NEB) at 16ºC for 4 h. A PCR was performed using 1 µl of ligation products and 1 µM of each divergent primers #59 and #60 (See Table 3.1). Inverse PCR reaction products of "2007" and "2011" time-points were resolved and visualized in a 1% agarose gel. DNA fragments were extracted of the gel with QIAquick Gel Extraction Kit (QIAGEN) and sequenced.

3.6 Protein techniques

Proteins were routinely analyzed by polyacrylamide gel electrophoresis in denaturing conditions (SDS-PAGE). Unless otherwise specified, SDS-PAGE gels consist in 12 % acrylamide/bisacrylamide 37.5:1 (Bio-Rad), 250 mM Tris-HCl pH 8.8 and 0.1% (w/v) SDS. Protein samples were dissolved in loading buffer (50 mM Tris-HCl pH 6.8, 2% SDS, 10%

(v/v) glycerol, 1% (v/v) β-mercaptoethanol, 12.5 mM EDTA, 0.005 % (w/v) bromophenol blue) and incubated at 90ºC for 5 min before loading. Gels were run at 100 mV in running buffer (25 mM Tris-base, 192 mM glycine, 0.1% SDS). Visualization was carried out staining the gel with staining solution (0.05% (w/v) Coomassie Brillant blue R-250 (Bio-Rad), methanol 50% (v/v) and acetic acid 10% (v/v)) and de-staining it with 10% methanol and 7% acetic acid solution. Precision Plus Protein Standards (Bio-Rad) was used as molecular weight marker. Protein concentration was determined in a Bradford assay using BSA (Sigma Aldrich) as a standard.

Protein electromobility shift assays, were performed in polyacrylamide native gels, (7.5% acrylamide/bisacrylamide 37.5:1 (Bio-Rad), 40mM Tris acetate, 1mM EDTA, 10% (v/v) glycerine, pH 8.3). Electrophoresis was performed at 16mA (80 min, 4ºC) in TAE buffer.

Finally, selected proteins were fluorescently labeled with Alexa Fluor 350 C5-maleimide (Life Technologies) following the manufacturer recommendations.

3.7 Protein purifications

3.7.1 Purification of sliding clamps and related enzymes

3.7.1.1 *E. coli* β

The DnaN gene of *E. coli* MG1655 was amplified by PCR from genomic DNA with oligonucleotides #1 and #2, cloned (NcoI-BamHI) in vector pET16b (Novagen), and sequenced. This plasmid was used to overexpress β in *E. coli* BL21(DE3). 15 g of cells (dry weight) were resuspended in buffer A (100 mM TrisCl, 1 M NaCl, 2 mM EDTA, 5% glycerine, 1 mM β-mercaptoethanol, 1 mM PMSF, pH 7.2) and processed four times with a French press (Thermo Fisher Corporation). All purification procedures took place at 4ºC. The lysed cells were diluted to 150 ml and centrifuged (15000 g, 30 min, 4ºC). The supernatant was treated with 0.2% polyethyleneimine and the precipitate removed by centrifugation. The supernatant was then brought up to 50% saturation (w/v) with $[NH_4]_2SO_4$. After centrifugation, the supernatant was further treated with $[NH_4]_2SO_4$ to 70% saturation (w/v) and centrifuged. The precipitate, containing β, was dissolved in 27 ml of buffer B (100 mM TrisCl, 100 mM NaCl, 2 mM EDTA, 5% glycerine, 1 mM β-mercaptoethanol, pH 7.2) and dialyzed against 2 l of this buffer for 12 h. This fraction (27 ml) was then applied on a 30 ml Q sepharose FF (GE Healthcare) ion-exchange chromatography column equilibrated in buffer B. Protein was eluted with buffer B and a NaCl gradient (320 ml, 0.1-1.0 M) and fractions (4 ml) collected. Fractions containing β (30

ml) were pooled and dialyzed against 1 l of buffer C (50 mM TrisCl pH 7.2) + 1 M $[NH_4]_2SO_4$. This fraction (250 mg) was then applied on a 20 ml Phenyl Sepharose 6 FF (GE Healthcare) chromatography column equilibrated in Buffer C. Hydrophobic interaction chromatography was performed in buffer C with a double gradient (100 ml) of $[NH_4]_2SO_4$ (1.0-0 M) and ethylene glycol (0-60%). Fractions containing β were pooled and dialyzed against 2 l of Buffer D (50 mM TrisCl, 50 mM NaCl, 1 mM EDTA, 10% glycerine, 1 mM DTT, pH 8). This fraction (4 mg ml^{-1}) was aliquoted and stored at -80ºC. In Fig. 4.5, SDS-PAGE gels of different steps of the purification are shown.

3.7.1.2 *Acidiphilium.sp* PM β

Acidiphilium sp. PM β was amplified by PCR using oligonucleotides #3 and #4, cloned into pET16b (Novagen) as a NcoI/BamHI fragment, and overexpressed in *E. coli* BL21(DE3). 10 g of cells (dry weight) were resuspended in Buffer E (100 mM Tris-HCl, 1M NaCl, 2 mM EDTA, 10% glycerine, 1 mM β -mercaptoethanol, 1 mM PMSF, pH 8) and processed four times with a French press at 4 °C. The lysed cells were diluted to 150 ml and centrifuged (15000 g, 30 m, 4 °C). β was mostly found in pelleted fraction in the form of inclusion bodies. The pellet was washed three times with 20 ml of buffer E + 1% Triton X-100 (Bio-Rad) to remove cell membranes and residual membrane proteins, and then centrifuged (15000 g, 30 m, 4 °C). The inclusion body pellet was then washed three times with 20 ml of buffer E to remove residual Triton X-100. In a subsequent step, inclusion bodies were solubilized in 5 mL of buffer F (100 mM Tris-HCl, 100 mM NaCl, 2 mM EDTA, 10% glycerine, 1 mM β -mercaptoethanol, 1 mM PMSF, pH 8) + 6M guanidine·HCl and incubated (25 °C, 15 m). Insoluble material was removed by centrifugation 15 minutes at 15000 g. The solubilised inclusion bodies were slowly drop by drop diluted in 400 ml of buffer F with constant stirring at 4 °C and allowed to refold for 1 hour. The solution was centrifuged 20 minutes at 15000 g to remove insoluble material. This clarified solution was applied on a 30 ml Q sepharose FF (GE Healthcare) ion-exchange chromatography column equilibrated in buffer F. Protein was eluted with Buffer F over a NaCl gradient (120 ml, 0.1–1.0 M). Fractions containing β were pooled and dialyzed against 2 l of Buffer G (50 mM Tris-HCl, 50 mM NaCl, 1 mM EDTA, 10% glycerine, 1 mM DTT, pH 8). This fraction (1.5 mg ml^{-1}) was aliquoted and stored at -80°C.

3.7.1.3 *Leptospirillum ferooxidans* β

To amplify the β gene from *Leptospirillum ferrooxidans* by PCR we used oligos #5 and #6. These oligos introduced a sequence encoding a FLAG epitope at the N-terminus of the gene. The PCR product was cloned as a BglII/NheI fragment into pET11a (Novagen). *L. ferrooxidans* β formed inclusion bodies when overexpressed and was purified following the same protocol as for *Acidiphilium sp. PM* β. The final yield was 1.3 mg ml^{-1}

3.7.1.4 *Methanosarcina barkeri* PCNA

Genomic DNA from *Methanosarcina barkeri Fusaro* (DSM 804) was obtained from DSMZ (Braunschweig, Germany) and used to amplify the gene encoding PCNA by PCR using the oligonucleotides #7 and #8. The PCNA gene was cloned (NcoI-BamHI) into vector pET16b and the protein overexpressed in *E. coli* BL21(DE3). Purification of PCNA followed the same protocol as for *E. coli* β with the following differences: a) $[NH_4]_2SO_4$ was raised to 40% saturation to remove contaminants and then to 60% to precipitate PCNA; b) the NaCl gradient on the Q sepharose FF chromatography column was from 0.2 to 1.0 M.

3.7.1.5 Human PCNA

hPCNA was cloned into vector pET16b using oligos #9 and #10 and overexpressed in *E.coli* BL21 (DE3) with 1mM IPTG and 12 g of dry cells were used in subsequent steps. Purification resembles the same protocol followed for *E. coli* β and *M barkeri* PCNA purification. Differential precipitation with $[NH_4]_2SO_4$ was carried out first at 45% saturation, and then raised to 70% saturation to precipitate hPCNA. Q sepharose FF chromatography first and Phenyl Sepharose FF chromatography next, were performed using same buffers and elution gradients as in *E. coli* β purification. The final purification yield of hPCNA was 3 mg ml^{-1}.

3.7.1.6 *E. coli* γ-complex

The processivity clamp loader γ-complex of *E. coli,* consists in five different kind of subunits γ, δ, δ´, χ and ψ with a stoichiometry of γ_3: δ: δ´: χ: ψ (Pritchard *et al.* 2000). Since $\gamma_3\delta\delta'$ (200 Kda) is sufficient to load β onto DNA (Onrust et al 1991), we purified this complex. γ-complex purification generally involves purifying their monomers separately, reconstitution of the complex in vitro and a further purification step of the complete complex from un-reactive monomers (Onrust *et al.* 1995). Here, we develop a one step $\gamma_3\delta\delta'$purification and assembly method.

DnaX, holA and holB genes of *E. coli* MG1655 were amplified by PCR from genomic DNA with oligonucleotides #11 and #12; #13 and #14 and #15 and #16 respectively. DnaX was cloned in pET16b (Novagen) as a NcoI/BamHI fragment which introduced a poly-histidine sequence at the N-termini of γ. holA and hoB were both cloned as a NdeI/BamHI fragment in vector pET16b (Novagen). All constructions were verified by sequencing.

γ, δ and δ´ were overexpressed in three independent 2 l cultures of *E. coli* BL21(DE3) with 1 mM IPTG. Cultures were centrifuged (12000 g, 25 min at 4ºC), pellet resuspended in buffer H (50 mM TrisCl, 50 mM NaCl, 2 mM EDTA, 5% glycerine, 1 mM β-mercaptoethanol, 1 mM PMSF, pH 8) and processed four times with a French press (Thermo Fisher Corporation). Each of the lysed cell preparations were diluted to 30 ml in buffer H and centrifuged (15000 g, 30 min, 4ºC). The three supernatants, containing γ, δ and δ´, were pooled together, and $\gamma_3\delta\delta'$ was allowed to assemble *in vitro* for 15 min at 25ºC (Fig 3.3 A). The crude protein extract was brought up to 40% saturation with $[NH_4]_2SO_4$ and centrifuged. The precipitate, containing γδδ´ (Fig 3.3 B), was dissolved up to 40 ml of buffer H and dialyzed against 2 l of the same buffer overnight at 4ºC. This fraction was applied on a 25 ml Q sepharose FF (GE Healthcare) ion-exchange chromatography column equilibrated in Buffer H. γ-complex was eluted with buffer H and a NaCl gradient (160 ml, 50 mM-1 M) and fractions of 4 ml collected (Fig 3.3 C). Fractions were γ, δ and δ´ co-eluted were pooled and dialyzed against 2 l of buffer I (50 mM potassium phosphate, 300 mM NaCl, 10% glycerine, 1 mM β -mercaptoethanol, 10 mM imidazole pH 8). This fraction was applied on a 10 ml Ni Sepharose 6 FF (GE Healthcare) ion-exchange chromatography column equilibrated in Buffer I. Highly pure γ-complex was eluted with buffer I + 250 mM imidazole (Fig 3.3 D). Fractions containing γ-complex (1 mg ml^{-1}) were pooled, dialyzed in PBS buffer (140 mM NaCl, 2.7 mM KCl, 10 mM Na2HPO4, 1.8 mM KH2PO4, pH 7.3) and stored at -80 ºC (Fig 3.3 E)

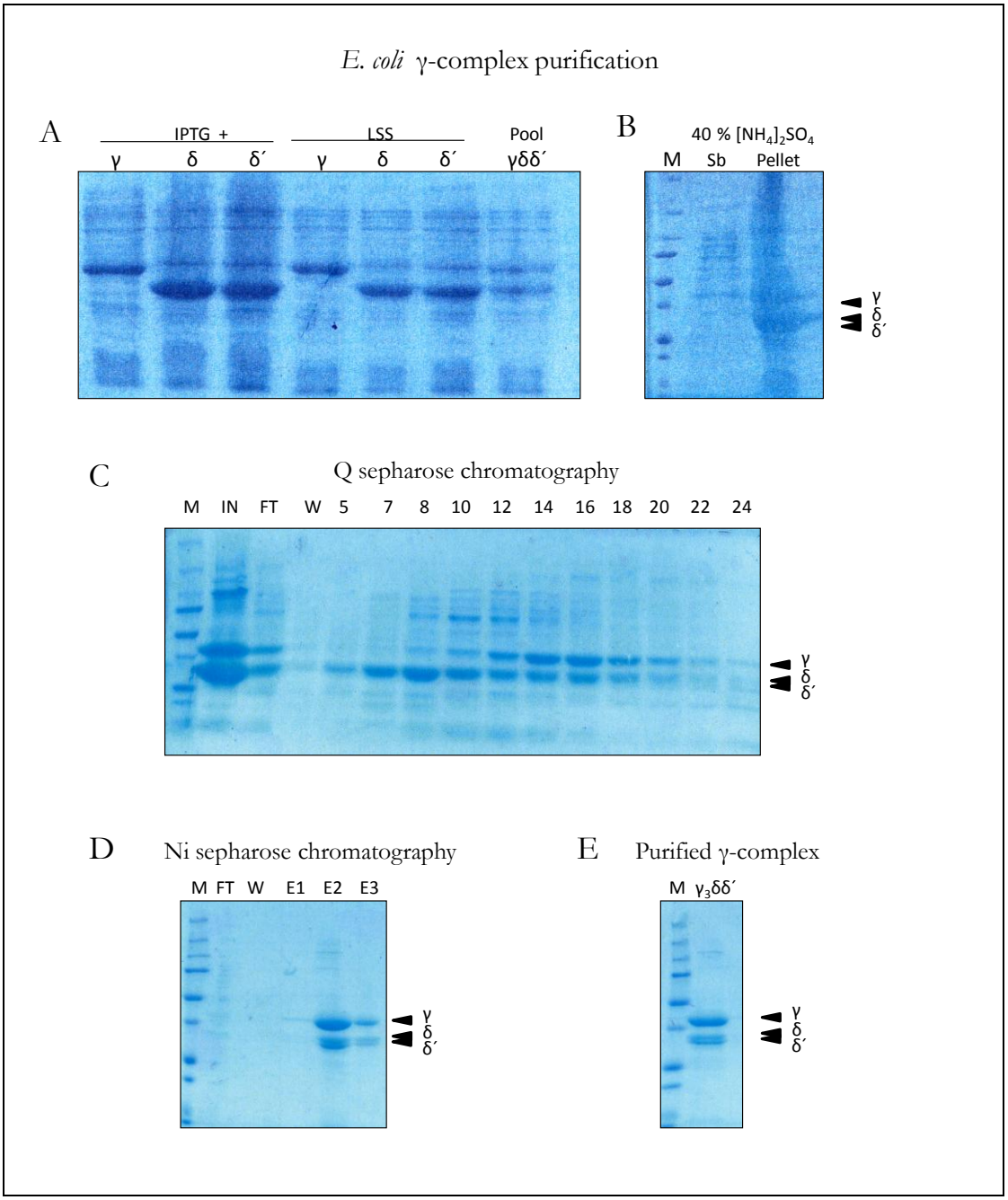


Figure 3.3. *E. coli* γ-complex purification.
A) Coomassie stained SDS-PAGE gel that represents protein extracts of overexpressed γ, δ and δ´ (lanes 1, 2 and 3 respectively). Low speed supernatants (LSS) were pooled together (lane 7). B) γ-complex is present in the pellet (lane 3) of a differential precipitation at 40% Ammonium Sulfate saturation. C) γ-complex was applied on a Q sepharose FF (GE Healthcare) ion-exchange chromatography column (IN) and eluted in a NaCl gradient. Fractions were γ, δ and δ´ co-eluted were pooled together and loaded on a Ni sepharose FF (GE Healthcare) ion-exchange chromatography column. D) Flow-through (FT), wash (W), first, second and third elution (E1, E2 and E3 respectively) of the Ni sepharose column. E) Eluted fractions containing γ-complex were pooled and dialyzed in storage buffer. Precision Plus Protein All Blue Standards (Bio-Rad) was used as molecular weight marker (M).

3.7.2 Purification of fusion proteins

3.7.2.1 GST-Pol IV^{LF}

The C-terminal domain of *E. coli* DNA polymerase IV (residues V243-L351) was amplified by PCR using genomic DNA from strain *E. coli* MG1655 and oligonucleotides #17 and #18. The PCR product was cloned (BamHI-XhoI) in pGEX-4T-3 (GE Healthcare). GST-Pol IV^{LF} (335 amino acids, 39 kDa) was overexpressed in 2 l cultures of *E. coli* BL21(DE3) with 1 mM IPTG, and cells lysed in a French press. The lysed cells were diluted to 40 ml in PBS buffer and centrifuged (15000 g, 30 min, 4ºC). GST-Pol IV^{LF} was soluble and present in the low speed supernatant. This fraction was applied on a Glutathione Sepharose 4B (Amersham) and eluted in buffer J (50 mM Tris-HCl, 10 mM reduced glutathione, pH 8.0). Purification yield was 1.8 mg ml^{-1}.

3.7.2.2 GST- LINE-1^{Z}, PIP1 and PIP2 mutants

The Z domain of the human LINE-1^{Z} ORF2 protein was amplified by PCR using oligos #19 and #20. The PCR product was cloned (BamHI-XhoI) in pGEX-4T-3 (GE Healthcare). GST- LINE-1^{Z} PIP1 (Q408A) and GST- LINE-1^{Z} PIP2 (Y414A; Y415A) mutants were generated by site-directed mutagenesis using oligos #21 and #22 and #23 and #24, respectively. Purification of these fusion proteins were performed following the same protocol as for GST-Pol IV^{LF}.

3.7.3 Purification of transposases

3.7.3.1 *Acidiphilum sp.* IS*1634* transposases

The gene for the IS*1634* transposase (Tnp) from *Acidiphilium sp PM* was amplified by PCR from genomic DNA with oligonucleotides #25 and #26, cloned in vector pET16b, and sequenced. The 5A (Q562A; F566A) and Cn (A563L; A564S; A565A) Tnp mutants were created by site-directed mutagenesis of the wt plasmid with oligos #27 and #28; and #29 and #30 respectively. Wild-type (WT) and mutant Tnps were overexpressed in Rossetta2 (DE3) pLysS (Merck Millipore) with 1 mM IPTG overnight at 25 °C. After centrifugation, the cell paste was resuspended in Buffer K (50 mM Tris-HCl, 400 mM NaCl, 2 mM EDTA, 10% glycerine, 1 mM β -mercaptoethanol, 0.1% Triton X-100, pH 8) and lysed in a French press. Cells were diluted to 150 ml in the same buffer and centrifuged (15000 g, 30 min, 4 °C). All three transposases formed inclusion bodies. The pellet was washed three times with 20 ml of buffer K supplemented with 1% Triton X-100, and then washed another three times with 20 ml of buffer K. Afterwards, the pellet was dissolved for 30

minutes at 25 °C in 5 ml of buffer K + 6M guanidine–HCl, insoluble material was removed by centrifugation (30 m, 15000 g). This solution was diluted drop by drop in 400 ml of buffer K with constant stirring at 4 °C and allowed to refold for 1 h. Soluble protein was applied to a 15 ml Heparin Sepharose FF (GE Healthcare) ion-exchange chromatography column equilibrated in the same buffer. Transposases were eluted with Buffer K and a NaCl gradient (120 ml, 0.4–1.0 M), 1.5 ml fractions were collected and those of them containing transposase were dialyzed against 2 L of Buffer K. Tnp was obtained at concentrations of 0.8 mg ml^{-1} for Wt; 0.5 mg ml^{-1} for 5A, and 0.5 mg ml^{-1} for Cn.

3.7.3.2 Tn*5* transposases

The full-length wild-type transposase (Tnp^{Wt}) gene of *E. coli* Tn*5* transposon was amplified by PCR from IS*50*R using oligonucleotides #31 and #32 and cloned in vector pET16b (Novagen Inc.) as a NcoI/EcoRI fragment which introduced a N-terminal poly-histidine sequence. Single amino acid mutants L363A (oligos #37 and #38) and L366F (oligos #39 and #40), and the 20 amino acids C-terminal truncation ($Tnp^{C\Delta 20}$) (oligos #41 and #42) were created by site-directed mutagenesis of the previous plasmid. The C-terminal 108 amino acid truncation ($Tnp^{C\Delta 108}$) was created by digestion and re-ligation of the wt plasmid with HindIII restriction enzyme. The N-terminal 7 amino acids deletion ($Tnp^{N\Delta 7}$), and the Inhibitory protein (Inh) were amplified by PCR from Tn*5* DNA using oligonucleotides #33 and #34, and #35 and #36 respectively and cloned (NcoI/EcoRI) in pET16b. Further steps of purification for all these proteins are essentially the same than steps followed for Tnp^{Wt} purification.

pET16b-Tnp^{Wt} plasmid was used to overexpress Tnp^{Wt} in 3 litres culture of *E. coli* BL21 (DE3) with 1mM IPTG at 30ºC. Cells were harvested by centrifugation (12000 g, 25 min at 4ºC), resuspended in buffer L (20 mM Tris-HCl, 400 mM NaCl, 20% glycerine, 0.1% Triton X-100, 2 mM EDTA, 1 mM β-mercaptoethanol, 1 mM PMSF, pH 7,2) and lysed four times through a French press. The lysate was diluted up to 100 ml in buffer L and centrifuged (15000 g, 30 min, 4 °C). Tnp^{Wt} remained mainly in the soluble fraction. To this supernatant polyethyleneimine (PEI) was added to a final concentration of 0.2% and the precipitate removed by centrifugation. The supernatant was brought up to 45% saturation with $[NH_4]_2SO_4$. Tnp^{Wt} was present in the precipitate which was dissolved in 20 ml of buffer L and dialyzed in the same buffer overnight. Then this soluble protein preparation was applied to a 15 mL Heparin Sepharose FF (GE Healthcare) ion-exchange chromatography column equilibrated in Buffer L. Protein was eluted with 120 ml of buffer

L and a NaCl gradient (0.4 - 1M) and 1.5 ml fractions were collected. Eluted fractions were analysed by SDS-PAGE (Fig. 4.12) and those of them containing Tnp^{Wt} free of contaminants pooled and dialyzed in PBS buffer, pH 7.2.

3.8 Protein interaction techniques

3.8.1 Protein-protein pull-down assays

3.8.1.1 *Acidiphilium* IS*1634* Tnp pull-down assay

The binding assay to test the interaction of IS*1634* Tnp (Wt, 5A and CN) with β clamp of *Escherichia coli*, *Acidiphilium sp PM*. and *Leptospirillum ferrooxidans*, used Dynabeads M-280 Tosyl-activated (Invitrogen). The reactions (50 µl) contained 12 µM of Tnp covalently coupled to 1.2 mg of magnetic beads in binding buffer M (phosphate-buffered saline, 0.1% Tween-20, 0.1% BSA, pH 7.2). 5 µM of each β labeled with Alexa Fluor 350 C5-maleimide (Life Technologies) were incubated (15 m, 25 °C) with the Tnp-coated beads, and washed three times with binding buffer M to remove unbound β. Reactions were stopped with 1% SDS, subjected to SDS-PAGE electrophoresis and visualized on a UV transilluminator.

3.8.1.2 Tn*5* Tnp pull-down assay

The binding assay was performed coupling *E. coli* β clamp covalently to Dynabeads M-280 Tosylactivated (Invitrogen). Each reaction (50 µL) contained β 10 µM coupled to 1 mg of magnetic beads following manufacturer´s recommendation in binding buffer N (Phosphate-buffered saline, Tween-20 0.1%). 2,2 µM of each Tnp (Wt, Inh, NΔ7, CΔ20, CΔ108, L363A and L366F) were incubated for 10 m at 25ºC with the β coated magnetic beads and washed three times with binding buffer N. Reactions were stopped with 1% SDS and loaded in a SDS-PAGE gel and visualized by coomassie staining. pmols of retained transposase by β was analyzed measuring the densitometry of each transposase eluted band in the SDS-PAGE gel with ImageJ software (Schneider *et al.* 2012).

3.8.1.3 GST- LINE-1^Z pull-down assay

The binding assay to test the interaction of GST- LINE-1^Z and its mutants PIP1 and PIP2 with human PCNA, used Glutathione magnetic beads (Thermo Fisher Scientific). 10.5 µM of each fusion protein was couple to 1 mg of magnetic beads in 50 ul of binding buffer O (PBS, BSA 0.1%, Tween-20 0.1%) for 15 min at 25ºC and subsequently washed with the same buffer. Then, 3 µM of hPCNA labeled with Alexa Fluor 350 C5-maleimide (Life Technologies), was incubated with the GST- LINE-1^Z coated beads in 50 µl volume reaction in buffer O, for 15 min at 25ºC. GST was used as a negative control. After

extensive washes with buffer O, the reaction was stopped with 1% SDS and loaded in a SDS-PAGE gel and visualized on a transiluminator.

3.8.2 Peptide-protein pull-down assays

IS*1634* transposase derived peptides were obtained from ProteoGenix SAS. IS*1634* transposase derived peptides from *Acidiphilium sp. PM*, *Acidithiobacillus ferrivorans* and *Methanosarcina barkeri* were assayed for the interaction with β clamp or PCNA. 400 μM of biotinylated peptides were mixed in 50 μl in binding buffer P (50 mM Tris, 50 mM NaCl, 5% glycerine, 0.1% BSA) with 1 mg of Dynabeads M-270 Streptavidin (Invitrogen), incubated (30 m, 25 °C) and washed three times with the same buffer. 4 μM of labeled Ecoβ , Lfβ , Acβ or 3 μM labeled MbaPCNA were added to the beads in a reaction volume of 50 μl with binding buffer P, incubated (15 m, 25 °C), and washed three times with the same buffer to remove unbound proteins. Reactions were stopped with 1% SDS, loaded on a SDS-PAGE and analyzed on a UV transilluminator.

Assays using biotinylated peptides derived from *E. coli* transposases followed the same protocol as above with the following differences: a) 440 μM peptides were used and b) binding buffer used was (50mM TrisCl, 100mM NaCl, 5% glycerine, pH 7.5).

In the other hand, biotinylated peptides derived from LINE-1, were obtained from Thermo Fisher Scientific. They were assayed as previously described against 3.5 μM of fluorescently label hPCNA and using as binding buffer PBS, BSA 0.1%, Tween-20 0.1%.

3.8.3 Electrophoretic mobility shift assay

Transposase derived peptides were used to probe their interaction with *E. coli* β, by testing their ability to disrupt preformed PolIV^{LF}·Ecβ complexes. The competition assay was performed in 20 ml in Buffer Q (40mM Tris acetate, 1mM EDTA, 3% glycerine, 12% DMSO, pH 8.3) supplemented with labeled *E. coli* β (0.55 μM). GST-Pol IV^{LF} (3 μM) and the different peptides (100 μM) were added as indicated in each figure. Reactions were incubated (10min, 25 ºC) and loaded on a native gel (7.5% acrylamide/bis 37.5:1, 40mM Tris acetate, 1mM EDTA, 10% glycerine, pH 8.3). Electrophoresis (80min, 16mA) was performed at 4 ºC in TAE buffer. The reaction products were visualized on a UV transilluminator.

Peptide titration assays were performed following the same protocol described above, but raising the concentration of GST-Pol IV^{LF} up to 6 μM. Peptide concentrations are indicated in each figure.

3.8.4 Crosslinking assay

Reactions were performed in 20 µl in Phosphate-buffered saline (PBS) pH 7.2 and supplemented as indicated with *E. coli* β 4 µM, IS*1634* transposase 4 µM, $BS(PEG)_5$ crosslinker (Pierce Biotechnology) 0.5 mM and incubated 30 minutes at 25 °C. Crosslinking reaction were quenched by incubating them with 50mM Tris-HCl, pH 7.5 for 15 min at 25 °C. Products were analyzed on a 8% polyacrilamide SDS-PAGE electrophoresis and visualize by Coomassie staining.

3.8.5 Fast protein liquid chromatography (FPLC)

Gel filtration assay was used to study the interaction between hPCNA and GST- LINE-1^Z. 20 µM hPCNA and 15 µM GST- LINE-1^Z were mixed in PBS buffer in a 50 µl reaction volume and incubated 15 min at 25ºC. Then, the reaction was applied on a Superdex 200 (GE) size exclusion chromatography column coupled to a FPLC system (ÄKTA, GE). PBS was used as mobile phase, and fractions of 80 µl at a constant flow rate of 0.2 ml/min, were collected. Fractions were analyzed on a 12% SDS-PAGE.

3.9 Protein-DNA interaction technique

Tn*5* wt and mutants transposases (Tnp) were tested for the interaction with DNA of the Kanamycin resistant gene flanked by mosaic ends (5´P- CTGTCTCTTATACACATCT-3´) in the presence or absence of β. Interaction assay was performed in a 10 µl reaction volume of buffer R (50mM potassium acetate, 20mM Tris-acetate, 10mM Magnesium Acetate, 1mM DTT, 30ng of Herring DNA, pH 8) with 6nM DNA, a titration of Tnp (90 nM, 180 nM and 360 nM), and 600nM β as indicated. Reactions were incubated 10 min at 25ºC, subjected to 1% agarose electrophoresis in TAE Buffer (65mV, 90 min at 4ºC) stained with SYBR Green (Invitrogen) and analyzed on a transilluminator.

3.10 Bioinformatic survey

Bioinformatic analysis of transposase gene identification and orientation in sequenced bacterial genomes was performed by the Bioinformatic Unit of the Centre for Astrobiology (INTA-CSIC), Madrid (Spain).

3.10.1 Genomic Data Set and Computational Pipeline

File collections containing orientation and coordinates of protein coding genes (*.ptt), predicted protein sequences (*.faa), and chromosomal nucleotide sequences (*.fna) of

partially and completely sequenced prokaryotic genetic elements were downloaded from the bacterial section of the National Center for Biotechnology Information (NCBI) Genome database, on October 24, 2012, as well as a summary file containing a table that linked accession numbers, replicon type (chromosome, plasmid), and taxonomic name. A computational pipeline written in Perl allowed navigation across the whole collection of files and directed the execution of a number of public domain or in house developed applications to detect, classify, and count IS elements according to their orientation, as described in the following sections. The working, curated data set consisted of 2,074 completely sequenced, circular, bacterial chromosomes, out of which 1,806 contained at least one IS (harbored by 1,685 species or strains).

3.10.2 IS Detection and Classification

The collection of predicted proteins from the genomic data set (6055750 sequences) was aligned with HMMER 3.0 against the Pfam 26.0 database (pfam.sanger.ac.uk, last accessed March 19, 2014) of domain profiles (Punta *et al.* 2012), using domain-specific score thresholds to filter the hits. The output of HMMER was processed with a Perl script to reconstruct protein architectures using a positional competition strategy to assemble the predicted protein domains allowing no overlaps. IS-related proteins were identified by comparing the new annotations against a list of 286 architectures that were considered characteristic of proteins encoded by IS elements and that were composed by a restricted collection of Pfam domains (Table I.1 in Appendix I). The architecture list was generated by manually extracting IS-encoded protein descriptions from the Pfam database and characterizing the domain structure of IS encoded proteins from the ISfinder database (www-is.biotoul.fr, last accessed March 19, 2014) (Siguier al. 2006; Punta *et al.* 2012). We were able to identify 80443 IS-associated genes. Once IS-related proteins had been identified in the set of bacterial genomes, IS elements were predicted following a strategy, articulated in four steps, that took into account that ISs can be composed of several genes and that they can appear in chromosomes as tandem insertions, making difficult the definition of their boundaries. In the first step, clusters of consecutive IS genes (separated by intergenic distance ≤500 bp) were identified in all genomes to calculate distance distributions for all possible pairs of ISrelated gene types (as defined by the architecture of the corresponding gene products).

In the second step, cluster detection was repeated, this time restricting the allowed intergenic distances to gene pair-specific distance ranges, deduced from the previous step

(mean±2 SD). Clusters detected in this step had ten genes at most. In the third step, the resulting collection of clusters was used to manually derive a list of 209 clusters that were accepted as representatives of the genetic organization of complete IS elements, on the basis of their correspondence to described IS structures (Table S2 online[1]), abundance (assuming that highly abundant and distributed clusters should correspond to complete IS elements), and length (in terms of number of genes). Each of these clusters was classified as belonging to a particular IS family. Accepted clusters had three genes at most. In the fourth step, each IS gene cluster detected in the second step was decomposed into all possible collections of nonoverlapping accepted subclusters to identify the collection that maximized the length of subclusters. Each subcluster from the optimal collection was then assigned to a particular IS family following the correspondences established in the list of accepted clusters (Table S2 online). A total of 57515 subclusters were detected, each of them representing a complete IS, that comprised 69438 (86%) of the IS-related genes. The remaining genes could correspond to chimeric or degenerated IS elements or be composed of protein architectures with ambiguous correspondence to IS families. To validate the IS detection and classification system of our computational pipeline, we performed two tests.

For the first test, we estimated transposase-related gene prediction recall relative to the annotations compiled in PTT files. The total number of genes, in the set of 2074 completely sequenced, circular, bacterial chromosomes, whose annotation in PTT files contained the string "transpos" was 65230. A total of 55800 of them were identified as IS-related genes by the pipeline, which implies an 85% recall rate. Thirty-four percent of the recovered genes had been annotated simply as "transposase" in PTT files. The pipeline identified 24643 additional IS encoded genes. For the second test, we determined IS family classification accuracy by comparison against the complete set of annotated prokaryotic chromosomes available, in April 2013, from the genomic component of ISfinder (ISbrowser) (Kichenaradja *et al.* 2010). The comparison involved 866 genes, coming from 33 chromosomes, that had been described as constitutive of IS elements by both ISbrowser and by our computational pipeline. The fraction of genes, considered globally, in which IS family affiliations coincided was 88%. The fraction of genes, by IS family, in which IS family affiliations coincided had average and median values of 79% and 100%, respectively.

[1] Table S2 online
http://gbe.oxfordjournals.org/content/suppl/2014/03/08/evu052.DC1/GBE_Gomez_SI_revised.pdf

3.10.3 Test on the orientation distribution of IS elements in chromosomes

The orientation of each IS element was defined by the orientation of its transposase gene relative to the local GC skew sign. GC skew [(G-C)/(G+C)] reflects an asymmetric nucleotide composition of the leading and lagging strands in Bacteria. The mechanism by which this asymmetry is created is unclear, but it could be related to the mutational or selective pressures on each DNA strand (Rocha *et al.* 2006). In genomes that have not suffered recent rearrangements, the two replicores have a different GC skew sign. GC skew was taken as a proxy for the direction of movement of the replication fork to correct for the effect of recent genome rearrangements. For each genome, GC skew was calculated with a Perl script over nonoverlapping 3001-bp-long genome segments. Then, a second script identified genome blocks with a minimal length of 10000 bp that were composed of consecutive segments having the same GC skew sign but allowing the inclusion of segments with the opposite sign if they were shorter than 10000 bp. Two blocks with different GC skew sign, corresponding to the replicores defined by the positions of the origin of replication and the termination site, were identified in 60% of the chromosomes.

The occurrence of multiple blocks in the remaining genomes can be explained in part as consequence of recent genomic rearrangements. IS element orientation relative to local sign of GC skew was defined as same (s) when the coding strand of the transposase gene had the same sign as that of the container GC skew block and anti (a) when the signs were different. The output of the pipeline consisted of a pair of orientation counts (s, a), for each genome and for each IS family, describing the number of IS elements presenting either of the two possible orientations.

To test whether IS elements were distributed randomly in the pair of orientation classes, counts were contrasted against a random binomial distribution with P=0.5 and a two-tailed P value was calculated (Table S3 online[2]).

3.10.4 Test on the orientation distribution of IS elements at phylum level

We then set out to determine whether IS families showed a bias at Phylum level by combining information from the chromosomes in each Phylum. The P values obtained in the tests on the orientation distribution of IS elements at chromosomal level (cumulative binomial probabilities, calculated as described earlier) were taken as a measure of the lack

[2] Table S3 online
http://gbe.oxfordjournals.org/content/suppl/2014/03/08/evu052.DC1/GBE_Gomez_Table_S3.pdf

of randomness in IS orientation for individual species. To obtain a test statistic representing a combined measure of the lack of orientation randomness for each IS family at Phylum level, we obtained the product of the *P* values calculated on the corresponding chromosomes, following a strategy similar to that of Bailey and Gribskov (1998). To minimize database bias, we chose randomly only one chromosome per species (1215 chromosomes; see Table S3 online, chromosomes labeled red). Because the distribution of such statistic is unknown, statistic values were contrasted against distributions generated after 106 sample, IS family-specific Monte Carlo simulations that assumed the random orientation of IS elements (Besag and Clifford 1991).

Simulated data sets had the same IS distribution of the original data in terms of number of chromosomes, number of IS copies per chromosome, and number of IS copies per IS family. Left-tailed *P* values were calculated as the fraction of simulated samples whose value was equal or lower than the value of the statistic calculated on the original data.

3.10.5 Detection of β-binding motifs in *Escherichia coli* transposases

To search for the β-binding motif among *Escherichia coli* transposases, we downloaded a collection of 2578009 *E. coli* protein sequences from the NCBI database (July 9, 2012) from which we identified 53235 IS-related proteins after assembling Pfam-based architecture descriptions as described earlier. We then used BlastClust (Altschul *et al.* 1997) to generate a subset consisting of 10980 unique sequences. Both the nonredundant set of *E. coli* IS-encoded sequences and the collection of 80443-associated genes derived from the genomic analysis were analyzed with the application Fuzzpro of the EMBOSS package (Rice *et al.* 2000) to explore the occurrence of β-binding motifs (QLSLF and selected derivatives). Hits were filtered with a Perl script and then manually curated.

Results

4.1. IS orientation in genomes and its interaction with the replication machinery

Current advances in whole genome sequencing are enlightening the diversity, wide distribution and impact of genetic mobile elements in organism´s genomes. Remarkably, genes encoding transposases and related elements are the most prevalent genes in nature (Aziz et al 2010). Although highly autonomous, IS transposition could be regulated or linked to various host mechanisms. Interestingly, there are several evidences that connect replication with transposition in some IS elements, among others, IS*1* (Zerbib *et al.* 1985), IS*50* (Yin el al. 1988), IS*903* (Hu and Derbyshire 1998), Tn*7* (Wolkow *et al.* 1996; Parks *et al.* 2009), Mu (Nakai *et al.* 2001) or IS*608* (Ton-Hoang *et al.* 2010). Although transposition pathways often require host factors implicated in DNA replication (Curcio and Derbyshire 2003; Turlan *et al.* 2004; Parks *et al.* 2009; Jang *et al.* 2012), no general mechanism linking these two processes has been proposed.

4.1.1 Orientation biases of IS families in bacterial chromosomes

The interaction between chromosomal replication and other cellular processes shapes the structure and organization of bacterial chromosomes (Rocha 2008). ISs are a class of genetic elements which they could potentially be placed almost anywhere in the chromosome and in any orientation. However, if transposition is mechanistically linked to replication, the interplay between the two processes could be reflected in detectable chromosome-wide patterns. In order to investigate the possible relationship between transposition and replication we analyzed patterns of orientation of ISs in fully sequenced bacterial chromosomes.

First we aimed to identify and classify all IS in complete sequenced bacterial genomes. The collection of predicted proteins from the genomic data set was aligned against the Pfam 26.0 database of domain profiles to reconstruct each protein architecture. IS-related proteins were identified by comparing the new annotations against a list of 286 architectures that were considered characteristic of proteins encoded by IS elements and that were composed by a restricted collection of Pfam domains (Table I.1 in Appendix I).This architecture list was created by retrieving IS-related protein descriptions from Pfam database and the domain structure of IS-encoded proteins from the ISfinder database (Table I.2).

Second, the orientation of each characterized IS element was defined by the orientation of its transposase. IS orientation patterns were investigated for each chromosome and each IS family by scoring the number of IS elements having either

orientation relative to the sense of movement of the replication fork as defined by the local GC skew sign to take into consideration possible recent chromosomal rearrangements. The GC skew [(G-C)/(G+ C)] reflects an asymmetric nucleotide composition of the leading and lagging strands in most Bacteria. The mechanism by which this asymmetry is created is unclear, but it could be related to the mutational or selective pressures on each DNA strand (Rocha *et al.* 2006). IS element orientation relative to local sign of GC skew was defined as same when the coding strand of the transposase gene had the same sign as that of the container GC skew block and anti when the signs were different.

We analyzed the orientation of 57,515 ISs in 1,806 completely sequenced circular bacterial chromosomes (Table S3 online[3]). We further analyzed only those cases in which six or more copies of a given IS family were found per chromosome and, to avoid database redundancy, in only one strain for each bacterial species (Table S4 online[4]).

We observed 153 cases of significant orientation bias ($P<0.05$) of IS families in chromosomes. These could mostly be assigned to a subset of eight IS families for which there was a bias in a large proportion of chromosomes containing six or more copies of the IS. Thus, families IS*200* (32% of chromosomes), IS*200*/IS*605* (25%), IS*607* (35%), and IS*NCYa* (20%) tend to be significantly biased for orientation in favor of the sense of advance of the replication fork (i.e., leading strand) in many chromosomes, whereas families IS*91* (25%) and IS*L3* (16%) show consistent bias for orientation against the sense of movement of the replication fork (i.e., lagging strand). Families IS*5a* (11%) and IS*110* (9%) showed no clear trend: in some chromosomes, the bias was toward a location in the leading, whereas in others, it was toward location in the lagging strand. Mapping of the biased IS families on the chromosomes showed that most IS insertions were well distributed and likely to be the result of independent transposition events (Fig. 4.1).

[3] Table S3 online:
http://gbe.oxfordjournals.org/content/suppl/2014/03/08/evu052.DC1/GBE_Gomez_Table_S3.pdf
[4] Table S4 online:
http://gbe.oxfordjournals.org/content/suppl/2014/03/08/evu052.DC1/GBE_Gomez_Table_S4.pdf

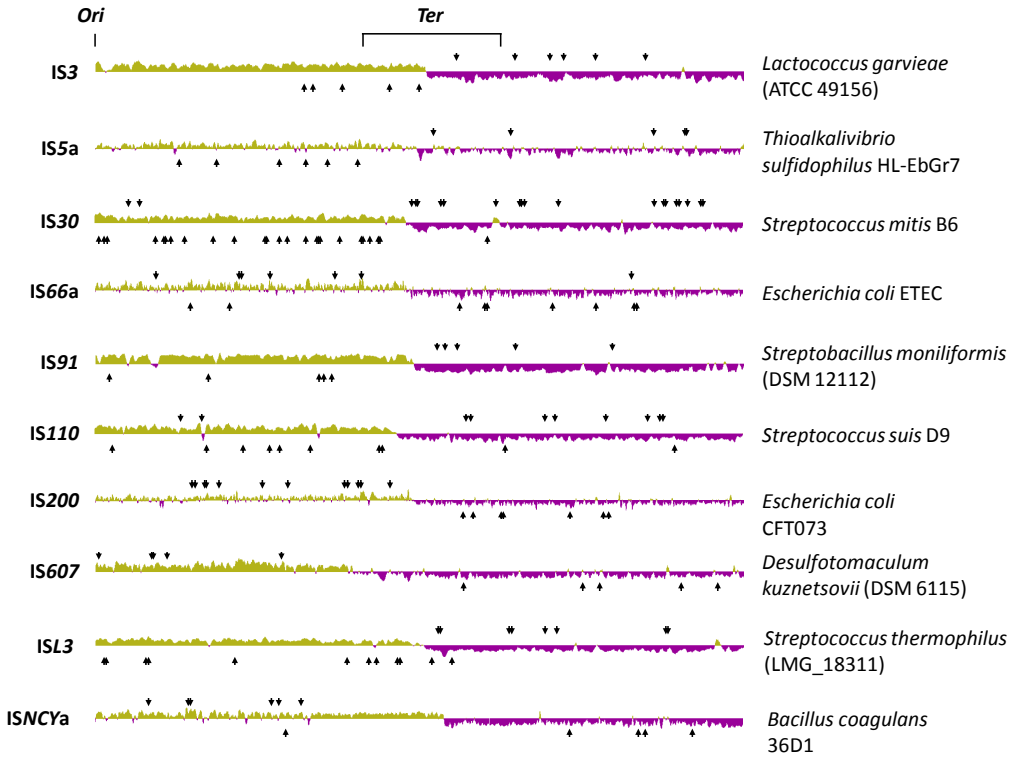


Figure 4.1. Graphic representation of IS orientation biases in bacteria.
Representative examples for ten IS families biased for their orientation in bacterial chromosomes. The trace represents the GC skew drawn using the program Artemis of the Sanger Institute (Rutherford *et al.* 2000) with a window size of 10 kb. Regions of positive GC skew (green) or negative GC skew (magenta) represent the two replichores. The arrows represent individual ISs inserted in the chromosome. For positive GC skew (green), an arrow pointing downward represents an IS oriented in the direction of movement of the replication fork, and an arrow pointing upward represents an IS oriented against the direction of movement of the replication fork. The opposite applies for regions of negative GC skew. For example, 42 of the 45 copies of IS*30* are oriented against the sense of advance of the replication fork in the chromosome of *Streptococcus mitis B6* (Binomial test, $p<10^{-9}$). Chromosomes are not drawn to scale.

Furthermore, we found numerous chromosomes in which two different IS families were significantly biased, either in the same or in opposite orientations (Fig. 4.2), in each case reflecting the pattern of IS bias specific for each family.

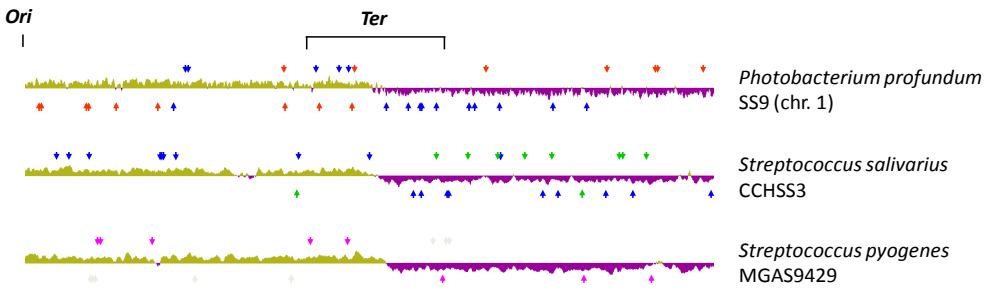


Figure 4.2. Graphic representation of IS orientation biases in bacteria (II).
Representative examples of chromosomes with two biased IS families of opposite orientations. For *Photobacterium profundum*, these are IS*200* (blue arrows) and IS*630* (red); for *Streptococcus salivarius*, IS*200* (blue) and IS*L3* (green); and for *S. pyogenes*, IS*3* (gray) and IS*As1* (magenta).

Many of the IS families analyzed (IS*1,* IS*3,* IS*4*a, IS*4*b, IS*4*c, IS*5*b, IS*5*c, IS*5*d, IS*6,* IS*30,* IS*66a,* IS*66*b, IS*256,* IS*481,* IS*630,* IS*701,* IS*982,* IS*1182,* IS*1380,* IS*1595,* IS*1634,* IS*As1,* IS*Azo13,* IS*NCYb,* Mu*,* Tn*3*, and Tn*7*) showed bias in few or no chromosomes. In some cases, the low number of chromosomes in which the IS was detected (IS*Azo13*) or the low number of ISs per chromosome (Mu, Tn*3*, and Tn*7*) precluded the detection of statistically significant differences. Also, it is likely that bias in IS families that had proliferated quickly and abundantly in chromosomes were easier to detect than in IS families which propagated slowly and in low numbers, as chromosomal reorganizations would blur any original orientation bias. We therefore sought to determine whether orientation biases could be generalized for higher taxonomic levels by considering the counts derived from groups of chromosomes in which a given IS family had been detected. Again, we included only one strain for each bacterial species (Table S3 online chromosomes marked in red). We calculated a new statistic for each IS family and compared its value against IS family-specific distributions derived from Monte Carlo simulations (see Materials and Methods). In addition to the families detected by observation of individual chromosomes, we found patterns of statistically significant biased orientation in ten IS families (Table 4.1). Thus*,* IS*66*a*,* IS*256*, and IS*As1* showed a tendency toward placement in the leading strand; IS*3,* IS*6,* IS*30,* IS*481*, and IS*NCYb* showed a tendency to be located in the lagging strand; and IS*630* and Tn*3* presented a mixed behavior. Further, we found that for many IS families, the biased patterns of orientation were, surprisingly, Phylum dependent. Thus, orientation bias is highly significant ($P<10^{-2}$) for ten IS families in Firmicutes but only for three IS families in Proteobacteria and two in Actinobacteria. No biased IS families were found in any other phyla (see Table I.3, for Bacteroidetes, Cyanobacteria, and Spirochaeta). Orientation bias in Firmicutes also was particularly strong, with P values $<10^{-5}$, for seven IS families. Analysis of the relative abundance of IS families in the different groups revealed that the orientation biases in Firmicutes did not arise from higher numbers of certain IS families in the chromosomes of these organisms (Table 4.1).

	Proteobacteria			Actinobacteria			Firmicutes		
	Chr.	IS	*Orient.*	Chr.	IS	*Orient.*	Chr.	IS	*Orient.*
IS1	33	1144	0.295	-	-	-	-	-	-
IS3	358	2601	0.273	100	968	0.380	159	1115	**5.95×10^{-4}**
IS4a	43	251	0.220	18	83	0.260	-	-	-
IS4b	17	82	0.116	-	-	-	8	21	0.511
IS4c	10	84	0.0903	-	-	-	-	-	-
IS5a	135	735	0.749	-	-	-	83	439	**5.00×10^{-6}**
IS5b	175	969	0.336	55	344	0.360	8	27	0.375
IS5c	51	192	0.684	-	-	-	-	-	-
IS5d	-	-	-	24	100	0.466	7	47	0.233
IS6	50	153	0.367	9	42	0.815	35	104	**1.21×10^{-3}**
IS21	203	787	0.0859	68	218	**0.0183**	161	509	0.380
IS30	80	289	0.241	53	186	0.200	91	500	**6.00×10^{-6}**
IS66a	135	640	0.0876	-	-	-	33	97	**0.0193**
IS66b	13	44	0.849	4	36	0.811	7	26	0.667
IS91	63	174	**6.53×10^{-3}**	-	-	-	11	26	0.138
IS110	260	1442	**6.10×10^{-5}**	73	439	0.345	112	721	**$<1 \times 10^{-6}$**
IS200	221	651	**$<1 \times 10^{-6}$**	-	-	-	89	270	**$<1 \times 10^{-6}$**
IS200/IS605	27	58	0.0901	11	23	0.399	38	90	**9.40×10^{-5}**
IS256	152	711	**0.0140**	79	578	0.510	97	564	**0.0161**
IS481	21	260	**0.0357**	43	147	0.493	-	-	-
IS607	-	-	-	8	30	**1.00×10^{-6}**	28	103	**$<1 \times 10^{-6}$**
IS630	145	1061	**0.0398**	39	252	0.379	25	120	0.565
IS701	47	234	0.223	33	168	0.118	6	26	0.204
IS982	30	285	0.813	-	-	-	24	145	0.653
IS1182	86	291	0.0545	26	82	0.294	79	386	0.312
IS1380	28	139	0.204	16	39	0.339	15	73	0.673
IS1595	95	333	0.340	6	26	0.157	7	33	0.519
IS1634	8	49	0.227	-	-		15	95	0.282
ISAs1	48	201	**0.0142**	14	249	0.756	8	54	**0.0138**
ISL3	91	336	0.343	56	333	0.270	84	496	**$<1 \times 10^{-6}$**
ISNCYa	68	221	0.139	-	-	-	53	203	**$<1 \times 10^{-6}$**
ISNCYb	-	-		10	72	**9.69×10^{-3}**	-	-	-
Tn3	70	138	**0.0186**	20	43	0.237	-	-	-
Tn7	46	61	0.0875	-	-	-	-	-	-

Table 4.1. Statistical significance for the non-random orientation in IS elements.
The table presents, for the groups Proteobacteria, Actinobacteria, and Firmicutes, the number of chromosomes in which a particular IS family was detected (Chr.), the total number of IS copies (IS) and the orientation test (Orient.), with P values representing the probability of obtaining, in 10^6 sample Monte Carlo simulations, a value as extreme as the one that was calculated from the observed data (see Methods for tests on the orientation distribution of IS elements at Phylum level). P values ($P<0.05$ in bold, $P<10^{-2}$ shaded gray) indicate an asymmetric distribution of IS elements in chromosomes in terms of their orientation relative to the local GC skew sign. IS families with less than 20 copies detected were omitted. See Table I.2 for nomenclature and IS classification scheme.

4.1.2 IS orientation biases are not generated by post-insertion selection

Our analysis of IS bias in chromosomes revealed that, for ten IS families, the Firmicutes showed a strong bias. Comparative genomics has shown that general gene orientation is nearly neutral for Proteobacteria or Actinobacteria but tends to be highly biased for Firmicutes and Tenericutes, where 78% of genes are co-oriented with the movement of the replication fork (Rocha 2002). It has been speculated that this phenomenon arises to prevent clashes between the DNA replication and transcription machineries, which can eventually lead to genomic instability, and substantial *in vitro* and *in vivo* experimental evidence seems to support this model (Wang *et al.* 2007; Srivatsan *et al.* 2010; Paul *et al.* 2013). However, it is not known why the orientation bias is especially strong in Firmicutes and Tenericutes and not in the other groups. Because ISs are typically very recent additions to chromosomes and do not encode essential or highly expressed genes, it is unlikely that they are subject to selection for any given orientation. However, to determine whether the observed IS orientation biases could just reflect a general preference for insertion in a specific orientation or the effect of selection, we determined global orientation for non-IS encoded and IS encoded genes in 1,727 bacterial chromosomes (Fig. 4.3). As previously observed (Rocha 2002), we detected a strong trend for non-IS genes to be placed in the leading strand (i.e., direction of movement of the replication fork) in Firmicutes and Tenericutes but not in other groups (Fig. 4.3 A). However, we found that IS-related genes (transposases and associated factors) show no orientation bias when considered globally in each chromosome, even in Firmicutes (Fig. 4.3 B). Further, there is no correlation between orientation bias for non-IS genes and IS genes in Firmicutes chromosomes (Fig. 4.3C), as it would have been expected if the processes that generate orientation biases were linked. This finding does not present a contradiction with our earlier result of many IS families showing strong orientation bias, as some IS families are consistently oriented in favor and others against replisome movement. In consequence, no net bias is found when considering IS genes globally in chromosomes. Indeed, numerous examples can be found of chromosomes in which different biased IS families show opposite orientation trends (Fig. 4.2).

Taken together, our data strongly suggest that IS orientation bias for the leading or the lagging strand is unlikely to be the result of selection for co-orientation of IS genes and fork movement but rather specific and intrinsic to the structure and mechanism of transposition of each IS family.

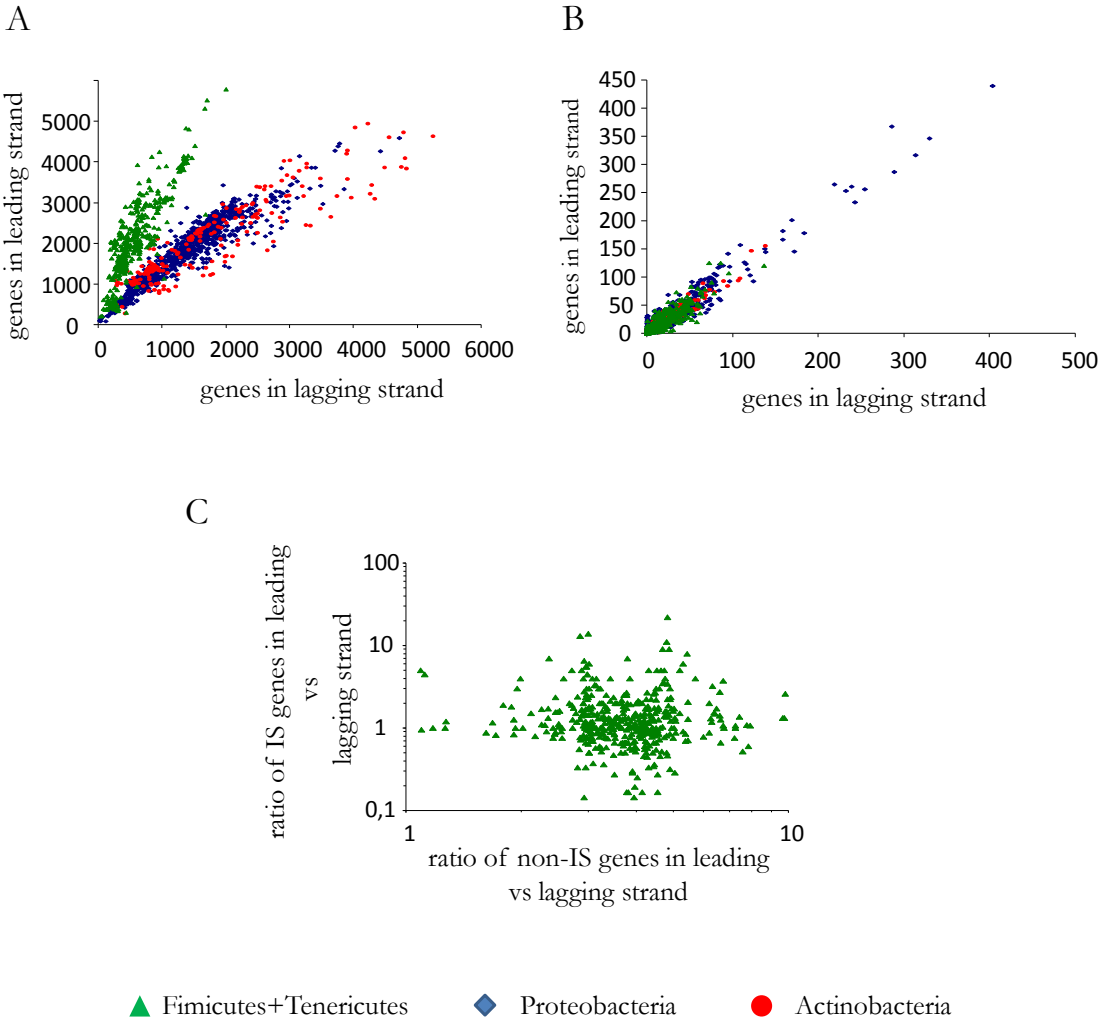


Figure 4.3. Global gene orientation for 1,727 circular bacterial chromosomes
A) General gene orientation excluding IS-related and RNA-encoding genes. The chromosomes of four major bacterial phyla are represented: Proteobacteria (blue diamonds, 1,009 chromosomes, 3,067,992 genes, leading/lagging strand ratio=1.07, R^2=0.89), Actinobacteria (red circles, 224 chromosomes, 837,682 genes, ratio=0.79, R^2=0.82), Firmicutes and Tenericutes (green triangles, 494 chromosomes, 1,287,236 genes, ratio=2.58, R^2=0.76). B) Orientation of IS-related genes (transposases and associated factors) in genomes for the Proteobacteria (44,504 transposase genes, ratio=1.05, R^2=0.94), Actinobacteria (9,202 transposase genes, ratio=1.00, R^2=0.92), and Firmicutes and Tenericutes (17,255 transposase genes, Ratio=0.97, R^2=0.79). C) Correlation between the ratios of IS-gene orientation and no-IS gene orientation is inexistent in Firmicutes.

4.1.3 Interaction of transposases with the β Sliding Clamp

We have shown the existence of a significant bias in the orientation of several families of insertion sequences relative to the movement of the replication fork at chromosomal level. Moreover, when we analyzed the orientation of IS families at Phylum level, we found out a stronger bias and in more IS families in Firmicutes than in other Phyla. We also discarded

that the IS orientation bias was originated by natural selection. Due to the antiparallel structure of the DNA, the replication machinery is associated to a number of asymmetric features related to the overall movement of the replisome and to differences in the synthesis of leading versus lagging strands. Replication asymmetries lead to various strand-specific differences and biases. For example, during DNA replication there is a transitory excess of ssDNA in the lagging strand which could explain the strong IS*200* bias for the lagging strand in most chromosomes both in Proteobacteria and Firmicutes, because its transposition mechanism requires ssDNA (Ton-Hoang *et al.* 2010). However, an excess of ssDNA would not account for the orientation bias found for other IS families in Firmicutes but absent in Proteobacteria or for IS families which transposition mechanisms are not dependent of ssDNA, therefore others replication mechanisms should also underlie transposition events.

Dissimilarities in the replication fork between Proteobacteria and Firmicutes may suggest what mechanism links replication with transposition and shed light on the observed stronger IS orientation bias in Firmicutes. Although mechanistically highly similar, the replication fork of Proteobacteria and Firmicutes differ in some relevant ways: In *E. coli*, only one DNA polymerase (DnaE) is responsible for processive synthesis, whereas in *Bacillus subtilis*, there are two (PolC and DnaE) (Sanders *et al.* 2010). Further, recent *in vivo* stoichiometric analyses using single molecule techniques of the *E. coli* replisome, have revealed that β sliding clamps remain bounded to the DNA during the lagging strand synthesis behind the replication fork for a prolonged period of time, reaching up to 20 clamps in each replisome (Moolman *et al.* 2014). A similar observation have been previously made for *B. subtilis*, where β accumulates during lagging strand synthesis but up to 200 clamps per fork, in a called "clamp zone" (Su'etsugu and Errington 2011). Moreover, in *B. subtilis* the chromosome is highly condensed and the left and right forks usually co-localize, so there are a vague physical separation between them (Migocki *et al.* 2004; Berkmen and Grossman 2006). In contrast, replichores in *E. coli* are positioned separately in opposite cell halves (Reyes-Lamothe *et al.*, 2008). Thus, in *B. subtilis* the local concentration of β in the replication focus and in the lagging strand may be higher than in *E. coli* replichores.

The β accumulation in "clamp zones" creates a molecular platform for β-binding enzymes related to DNA metabolism (Su'etsugu and Errington 2011; Moolman *et al.* 2014). Therefore, we reasoned that if transposases or related factors interacted with β, the difference in β amounts associated with lagging strand synthesis among Proteobacteria and

Firmicutes, could account for the observed biases in IS orientation between both Phyla. This, along with the previous described evidence that TnsE, a Tn*7*-encoded factor that targets transposition preferentially to replicating conjugative plasmids, interacts with the β sliding clamp (Parks *et al.* 2009), place β as a potential keystone in the interconnection between transposition and replication.

Within the replication fork, the β sliding clamp interacts with a large number of enzymes, which share a short and poorly conserved binding motif (consensus: Q-L-S-L-F/L, Q_1, and L_4 being the most strongly conserved residues) (Dalrymple *et al.* 2001). We therefore searched transposases and accessory factors for this specific sequence, focusing on *E. coli* because of the large number of sequenced genomes and because the interactions of β in this model organism have been extensively analyzed (López de Saro *et al.* 2003). In contrast to *E. coli,* no ISs have been detected in the *B. subtilis 168* chromosome. β-binding motif are often located on regions at the C-terminus of the protein (Dalrymple *et al.* 2001). We found several transposases of different IS families in *E. coli* with a putative β binding motif (Fig. 4.4) which is also conserved in other species (Alignments in Fig I.4 in Appendix I).

Pol IV (NP_414766)	VGLHVTLLDPQMER**Q**L**L**VGL	(C-ter)
IS*5*a (AAB53644)	QIQGVAENDN**Q**L**L**AM**L**FTLAN	(C-ter)
IS*30* (NP_415922)	YFPKKTCLAQYT**Q**HE**L**DLVA	(C-ter)
IS*66* TnpB (YP_424826)	RDGKVHLTPA**Q**L**L**SM**L**LEGIN	(C-ter)
IS*66* TnpC a (YP_003235004)	SEQAEALRQKDQ**Q**L**L**SL**L**VEET	(N-ter)
IS*66* TnpC b (ZP_07592975)	RFGKKCESLAGM**Q**R**L**SL**L**FEED	(N-ter)
IS*91* (ACO24927)	ERAPPLTPSLFDPS**Q**SR**L**L**F**D	(C-ter)
IS*200* (1) (ZP_03029803)	YARYQEKMEQTHEQ**Q**ME**L**L**L**E	(C-ter)
IS*200* (2) (NP_752024)	YIKHQLEEDKMGE**Q**L**L**SI**L**PYP	(C-ter)
IS*1380* (YP_003829282)	VLKPEKERA**Q**L**L**SL**L**EGSEYD	(internal)
IS*L3* (ZP_07122173)	MCEKEPELKIAQ**Q**L**L**VL**L**EFYR	(C-ter)
IS*NCYa* (EIL58166)	DVAEMANLPLAEID**K**V**L**INL**L**I	(C-ter)
Tn*3* (YP_001816608)	QKGTGATEIAH**Q**L**L**SI**L**ARSTV	(C-ter)
Tn*7* TnsC (EIL57895)	GPESEAYDRFK**Q**AG**L**I**L**LDLR	(C-ter)

Figure 4.4. List of peptides used in the biochemical assays aligned at the β -binding motif. Peptides derived from transposases found in *Escherichia coli* chromosomes and used in biochemical experiments are listed. Residues corresponding to the consensus β motif are in bold type and those mutated to alanine, underlined. Two peptides (a and b) were designed for different regions of two distinct TnpC proteins of IS*66*. Two homologous peptides (1 and 2) were designed corresponding to variants of IS*200* transposase. PolIV peptide was used as a positive control. NCBI accession numbers for protein sequences are shown.

Because transposases are frequently toxic and insoluble when overexpressed, we exploited the fact that β binding motifs are often located on highly flexible structures or in disordered regions at the C-terminus of the protein (Dalrymple *et al.* 2001; Bunting *et al.* 2003; López de Saro *et al.* 2003). Besides, unlike most protein–protein interactions, which implicate relatively large surface areas, interactions with β are mostly circumscribed to the interacting motif binding to a hydrophobic pocket on β (Georgescu *et al.* 2008). To investigate the biochemical interaction between transposases and β sliding clamp, we synthesized N-biotinylated peptides (20 aa) derived from sequences of transposases containing putative β -binding motifs (Fig. 4.4), as well as peptides in which Q_1 and L_4 had been changed to alanine, and assayed them for β binding. In addition, we purified *E. coli* β sliding clamp (see Methods and Fig. 4.5) and we fluorescently labeled it with Alexa Fluor 350 C5-maleimide (Life Technologies). Maleimide-labeling results in one label per β monomer at Cys-333 and do not alter its interaction with DNA polymerases or its activity in replication assays (Griep and McHenry 1988; López de Saro *et al.* 2003).

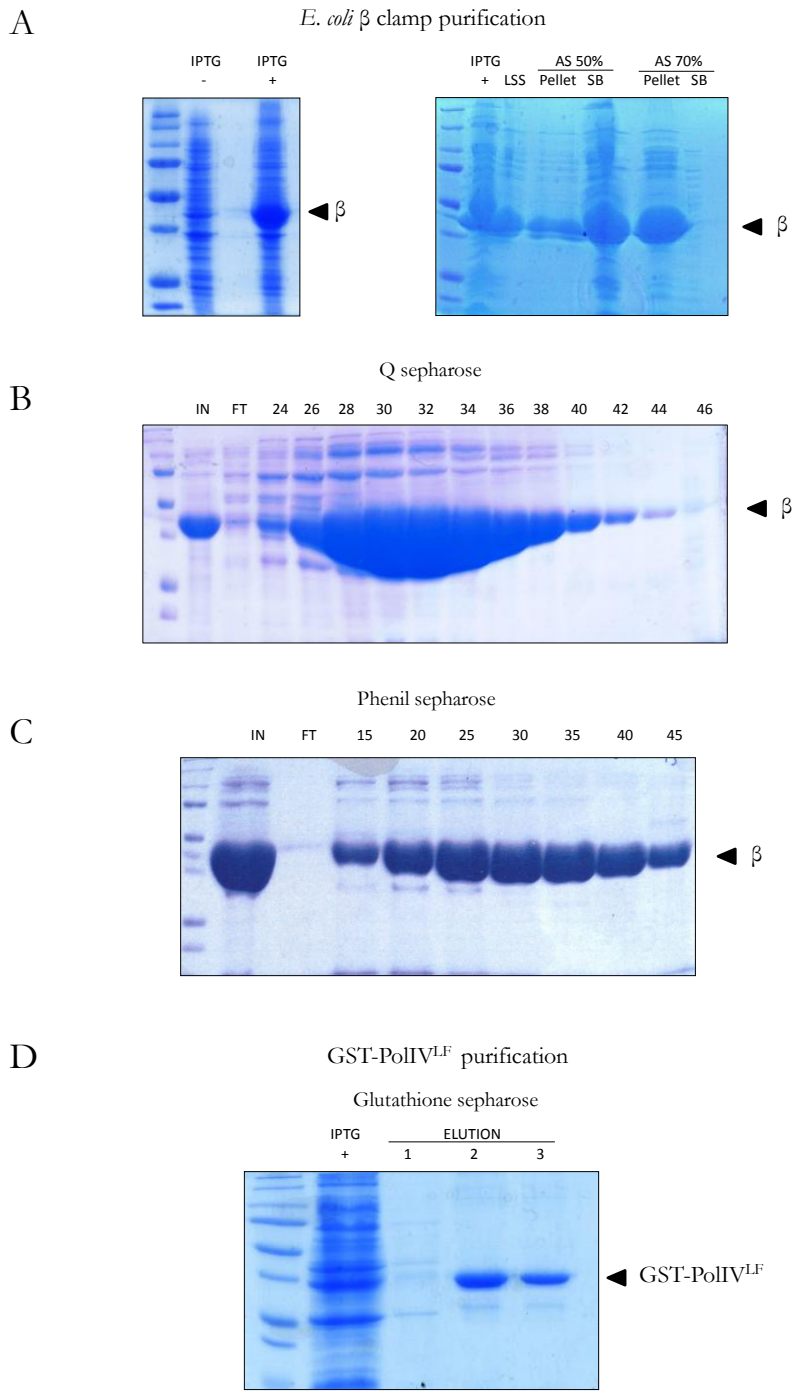


Figure 4.5. *E. coli* β clamp and GST-PolIV^{LF} purification.
A) β was overexpressed in *E.coli* BL21 with 1mM isopropyl β -D-1-thiogalactopyranoside (IPTG). Lane 2 and 3 in the left Coomassie stained SDS-PAGE gel represent a protein extract before (IPTG -) and after induction (IPTG +) respectively. In the right gel the low speed supernatant (LSS in lane 3) was subjected to a differential Ammonium Sulfate (AS) precipitation. At 50% Ammonium Sulfate saturation, β was mainly present in the soluble fraction (lane 5, SB). This fraction was brought to 70% AS saturation and β precipitated (lane 6, Pellet). B) This β containing fraction was dialyzed and applied on a Q sepharose FF (GE Healthcare) ion-exchange chromatography column (IN) and eluted in a NaCl gradient. Fractions were β was present were pooled together and dialyzed. C) β was loaded on a Phenyl Sepharose FF (GE Healthcare) chromatography column (IN) and eluted with a double gradient of AS (from 1.0 to 0 M) and

ethylene glycol (0-60%). Fractions containing β and free of contaminants were pooled, dialyzed and fluorescently labeled. D) GST-PolIV^{LF} was overexpressed in *E.coli* BL21 with 1mM IPTG (Lane 2). Soluble protein fraction was applied on a Glutathione Sepharose 4 Fast Flow (GE Healthcare) resin, washed and eluted three times with reduced glutathione (Lanes 3 - 5, Elution 1, 2 and 3). First lane in all SDS-PAGE gels is the molecular weight marker (Precision Plus Protein All Blue Standards (Bio-Rad))

To investigate the possible interaction of transposase-derived peptides with purified β, we followed two different approaches. First, we bound the peptides to streptavidin magnetic beads and tested their ability to bind and retain fluorescently labeled. After extensive washes, we found that peptides derived from transposases belonging to nine IS families (IS*5*a, IS*30*, IS*66*a, IS*91*, IS*200*, IS*1380*, IS*L3*, IS*NCY*a and Tn*7*) bind to β. However mutants where Q_1 and L_4 of the binding motif had been changed to alanine, do not retain β (Fig. 4.6).

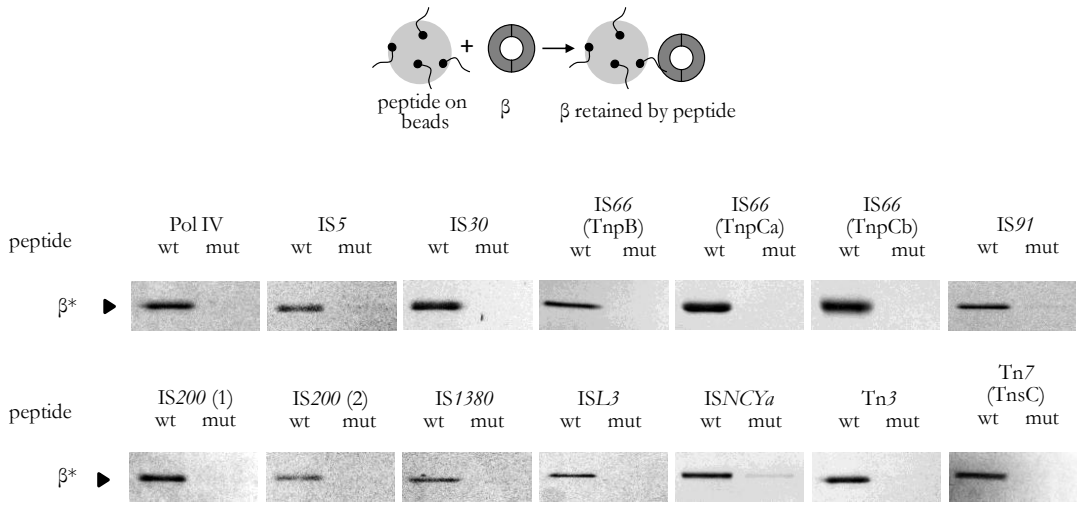


Figure 4.6. Transposase-derived peptides interact with *E. coli* β sliding clamp.
N-biotinylated peptides were coupled to streptavidin-coated paramagnetic beads and used to retain purified Alexa 350-labeled *E. coli* β. In each panel, the native (wt) and the mutated (mut) peptides were used as indicated.

Second, to determine the interaction site on the surface of β, we tested the peptides in a competitive, native gel mobility-shift assay, using the C-terminal domain ("little finger") of *E. coli* DNA polymerase IV (PolIV^{LF}). PolIV^{LF} (Fig. 4.5 D) binds strongly to a hydrophobic pocket on the surface of β that is also the binding site for the other four polymerases in *E. coli* and for various DNA repair factors (Bunting *et al.* 2003; López de Saro *et al.* 2003). Therefore, this competition assay would assure high specificity in the interaction, despite the relative simplicity of the consensus Q-L-S-L-F/L motif. The

peptides were tested for their ability to disrupt the β·GST-PolIV^{LF} complex by adding them in a molar excess to the reaction and then separating the products by a native gel electrophoresis. We found that all peptides that bound to β in the streptavidin-binding assay were also capable of binding β in competition with GST-PolIV^{LF} (Fig. 4.7), suggesting that they interact with β in the same fashion as other β ligands. The mutant peptide variants, again, were unable to compete with GST-PolIV^{LF}.

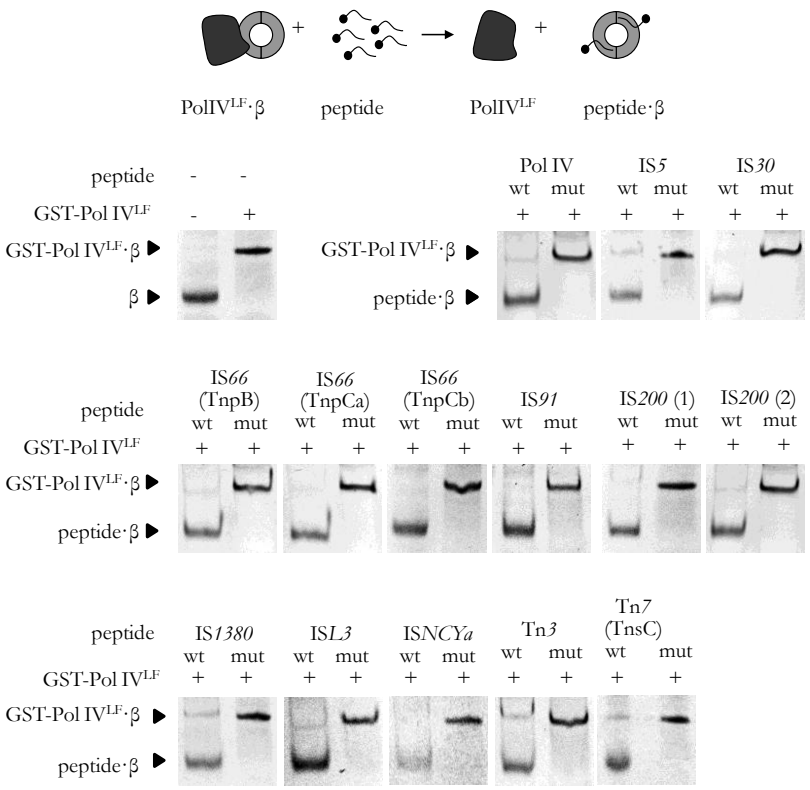
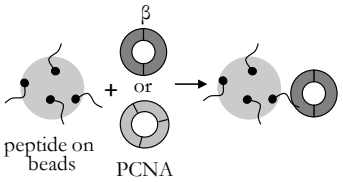


Figure 4.7. Transposase-derived peptides compete with PolIV^{LF} for binding *E. coli* β.
Fluorescently labeled β whose mobility was retarded in a native gel by interaction with GST-PolIV^{LF} was challenged with an excess of transposase-derived peptides, as indicated. While wild-type peptides (wt) disrupt the β·GST-PolIV^{LF} complex, mutants (mut) do not. The competition between transposase peptides and PolIV^{LF} indicates that transposases bind in the same hydrophobic pocket of β as PolIV^{LF} does.

The β-binding motifs found in these transposases are conserved in distant species (Alignments in Fig I.4), suggesting that interaction with β is widespread across IS of distant phylogenetic groups. Conservation is extensive to Archaeal ISs, but in this case, the PCNA motif is present (consensus Q-x-x-L/I-x-x-F-F) (Fig 4.8 and IS*200* alignments in Fig. I.4). In order to biochemically test whether transposase - sliding clamp interaction is maintained in Arhaea, we purified and labeled *Methanosarcina barkeri* PCNA and performed a pull down

assay using a peptide derived from *Mba*IS*200* transposase. Results in the left panel of Fig. 4.8 reveals that a *Mba*IS*200* peptide binds strongly to PCNA, meanwhile the mutated peptide in the Q_1 and F_7 residues of the canonical binding motif is not able to retain PCNA. Because the PCNA motif is a related variant of the β motif, we also tested if *Mba*IS*200* could interact with *E. coli* β. Indeed, *Mba*IS*200* interacts with *E. coli* β and point mutants of this peptide no longer bind to β (Fig. 4.8). Our results suggest that interaction with the replisome would not likely limit the transmission of archaeal ISs to bacterial chromosomes, and few mutations would be required to adapt a bacterial transposase to the archaeal replication machinery.



Mba*IS*200 (YP_307176) NQGNQEEKEAYK**Q**MK**I**ID**F**Q (C-ter)

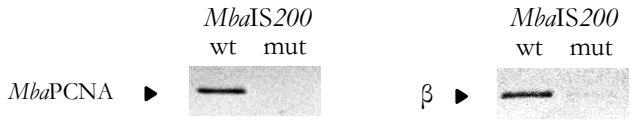


Figure 4.8. An archaeal transposase-derived peptide binds archaeal PCNA and *E. coli* β clamp.
N-biotinylated peptides derived from the C-terminus of *Methanosarcina barkeri* IS*200* were bound to streptavidin-coated magnetic beads and used to probe interaction with *M. barkeri* PCNA (left panel) or *E. coli* β (right panel). The *Mba*IS*200* peptide sequence used in the assay is listed. The PCNA consensus motif (bold) and the residues changed to alanine in the mutated peptide (underlined) are marked.

We have demonstrated that transposases from different IS families with distinct transposition mechanisms interact with β trough a conserved binding motif. Although the consensus sequence of the binding motif is Q-L-S-L-F, it accepts some degree of looseness. Hence, the broad collection of different binding motifs sequences in transposases, even within the same family (Fig. 4.4 and Alignments in Fig. I.4), raises the question of to what extent diverse binding motifs have distinct β affinity. We chose two transposase derived peptides belonging to the same IS family, IS*200* (1) and IS*200* (2), and we used a competition assay to analyze their β affinity. A titration of both peptides on a preformed complex β· $PolIV^{LF}$, and a quantification of the β·peptide complex, show that the IS*200* (2) peptide binds β with higher affinity than the IS*200* (1) (Fig 4.9 A and B).

These results point out that a protein with a binding motif which resembles to the canonical sequence (Q L S I in IS*200* (2)) binds stronger to β than another protein with farther sequence from the consensus motif (Q M E L in IS*200* (1)).

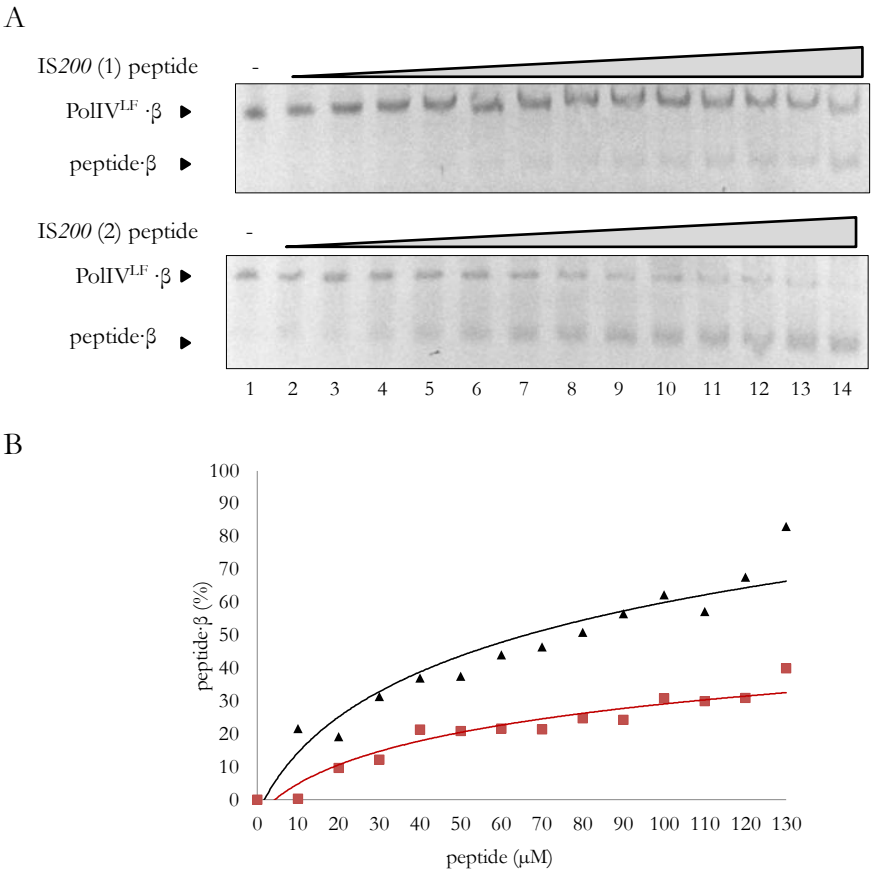


Figure 4.9. Two different binding motifs in transposases of IS*200* family, bind *E. coli* β clamp with distinct strength.
A) Competition assay of complexes of *E. coli* β bound to DNA polymerase IV (PolIV^{LF}) against a titration of IS*200* (1) (top panel) and IS*200* (2) (bottom panel) peptides. Concentration of peptides for lanes 1–14 ranging from 0 to 130, increase from 10 to 10 μM. B) Densitometry of complexes formed in A, were plotted as percentage of peptide·Ecβ bound. IS*200* (1) data is represented as red squares and IS*200* (2) as black triangles

Similarly, in members of the IS*66* family we found that the binding motif can be present in TnpB, TnpC or both, and in TnpC can be located upstream or downstream of the leuzine zipper domain (see Fig 5.1). We used the previously described competition assay to check out the relative affinity for β of three different IS*66*-derived peptides. We found that a peptide derived from IS*66* TnpC b has a remarkable higher affinity for the sliding clamp than the other peptides (Fig. 4.10 A and B). IS*66* TnpC b has a binding motif

sequence close to the consensus and it is located upstream of a leuzine zipper in a N-terminal region.

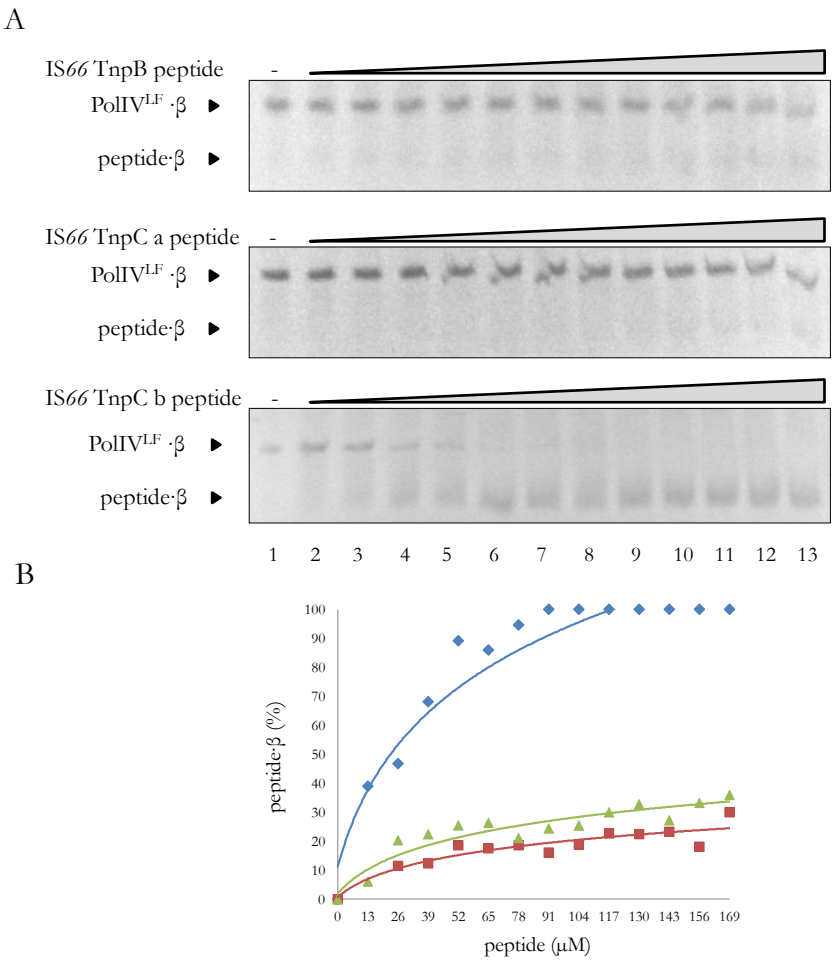


Figure 4.10. Different binding motifs located in transposases of IS*66* family, bind *E. coli* β clamp with distinct affinity.
A) Competition assay of complexes of *E. coli* β bound to DNA polymerase IV (PolIV^{LF}) against a titration (from 0 to 130 μM) of IS*66* TnpB (upper panel), IS*66* TnpC a (middle panel) and IS*66* TnpC b (lower panel) peptides. B) Percentages of bound peptide to β are represented as green triangles for IS*66* TnpB, red squares for of IS*66* TnpC a and blue diamonds for of IS*66* TnpC b.

The possible effect on transposition rates of a higher o lower β affinity by a transposase will be addressed in detail later (see section 4.3). Interestingly, the IS*200* transposase from which IS*200* (2) peptide is derived, is distributed among 41 different *E. coli* strains accounting for a total of 67 elements, meanwhile IS*200* (1) which has a lower β affinity is only present in two *E. coli* genomes in a single copy in each of them.

4.2 Tn*5*: Tnp interaction with β and characterization of novel hyperactive mutants

The interest on the study of the transposase of Tn*5* (Tnp) relies in its similarities with other members of the same superfamily, like HIV-1 integrase or MuA transposase. Although Tn*5* transposition has been extensively study, even in atomic detail (Davies *et al.* 1999), there are many knowledge gaps regarding critical steps in its mechanism. For instance, the mechanistic process behind DNA binding it is still not fully understood. The development of new mutants is also important for its practical use. In fact a hyperactive Tnp is commercially relevant because of its applications as a molecular biology tool.

DNA binding and synaptic complex formation (i.e. transpososome) are key and limiting steps in transposition. These tight self-regulated processes have been extensively study in Tn*5* and still today represent an intriguing paradox. Initial analysis of Tnp - DNA complexes, showed that Tnp alone was not able to bind outside ends of the transposon (OE) *in vitro*. Tnp - OE complexes detected contain heteromultimers of Tnp with either natural occurring Tnp truncations or Inh, generating presumably non productive complexes (De la Cruz et al 1993). Functional analysis using a collection of amino terminal truncated transposases, determined that the N-terminal domain of Tnp is required for the DNA binding (Weinreich et al 1994b). Interestingly, C-terminal truncated Tnp variants (e.g. lacking the last 107 amino acids) are able to bind DNA *in vitro* (Weinreich et al 1994b; York and Reznikoff 1996). The C-terminus of Tnp contains the dimerization domain necessary for the synaptic complex formation (Steiniger-White and Reznikoff 2000). Cross-linking and gel filtration experiments demonstrated that in solution Tnp is predominantly monomeric, while Inh (lacking N-terminal 55 amino acids) homodimerizes (Braam *et al.* 1999). Collectively, all these findings imply that in solution, the C and N-terminus establish intramolecular contacts, that limits Tnp from DNA binding nor dimerization. Thus, C-terminus blocks the DNA-binding domain (at the N-terminus) and the N-terminus precludes dimerization (at the C-terminus). Considering that Tnp has to dimerize and bind DNA for transpososome formation, it is required some sort of conformational change to dislocate both ends (Reznikoff 2008).

The principal obstacles in studying the mechanism of Tn*5* are that Tnp has no detectable activity *in vitro* and its limited ability to bind DNA. In fact, most genetic, biochemical and structural studies described to date, have been performed with a variable collection of Tnp hyperactive mutants. Since DNA binding is a limiting step in transposition, most of the random hyperactive mutants isolated, increase transposition rate by improving the Tnp-DNA binding (Zhou and Reznikoff 1997). A commonly used

mutation is the replacement of the methionine at position 56 with an alanine, which prevents the synthesis of the inhibitory protein (Wiegand *et al.* 1992). Other key mutation is L372P that is located into a C-terminal α-helix of the protein and increases transposition ratio by 10-fold by improving DNA binding (Weinreich *et al.* 1994c). This mutation likely breaks the helix and dislocates the auto-regulatory contacts established between the amino and carboxi temini. Remarkably, the commercially available Tn*5* transposase, gathers a combination of E54K, M56A, L372P and P242A, to reach a 100-fold increase of *in vivo* activity compared to the wild-type transposase (Tnp^{Wt}) (Reznikoff *et al.* 2006). It is also active promoting transposition *in vitro*, but since Tnp^{Wt} is not active *in vitro*, a total fold change increase could not be given.

Figure 4.11. X-ray co-crystal structure of Tn*5* synaptic complex.
The structure illustrates the crystal structure of a Tn*5* transposase bound to precleaved outside ends (OE) DNA, that is believe to be representative of the synaptic complex generated *in vivo* after the cleavage of the donor DNA. The structure is dimeric. Two Tnp interact with each other, and each Tnp bind to an OE DNA. To clarify the visualization, only one monomer (one Tnp bound to one OE) is represented. In blue ribbons are represented the 55 amino acids in the N-terminus of the protein that are absent in the inhibitor protein (Inh). In orange are represented the C-terminal 20 amino acids that are truncated in $Tnp^{\Delta C20}$. These amino acids are located in the α-helix implicated in the dimerization domain. Amino acids L363 (red) and L366 (magenta) are mutated to alanine and phenylalanine respectively in the hyperactive proteins described in this work. (PDB ID: 1MUS)

We have previously described, that transposases of different IS families interact with β sliding clamp. Here, we firstly study whether this interaction is also maintained between β and the transposase (Tnp) of the well-known Tn*5* transposon. Then, we

investigate how this interaction, could potentially influence on Tn*5* transposition mechanism *in vitro*. Finally, we seek point mutants that could increase the transposition activity of Tn*5* *in vivo*. Biochemical assays were performed using purified Tnp along with other truncation versions and selected point mutants of the enzyme (Fig 4.11 and Fig 4.12 for purifications). *E. coli* β was purified as previously described (Fig 4.5).

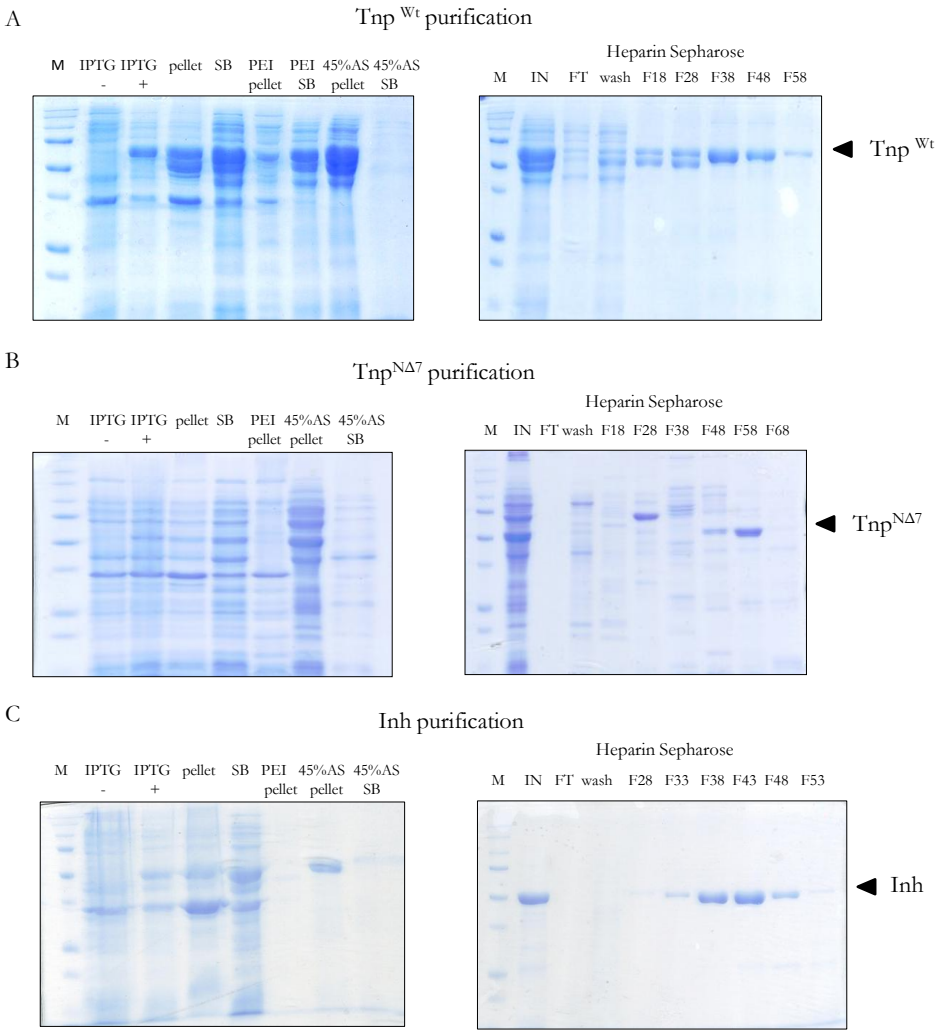


Figure 4.12. Tnp^{Wt}, Tnp^{NΔ7} and Inh purifications
Wild type transposase (Tnp^{Wt}) and its deletion variants, Tnp^{ΔC20}, Tnp^{ΔN7} and Inh, were overexpressed in *E. coli* Bl21 (DE3) with 1mM IPTG. Tnp were present in the soluble fraction (LSS). Then, crude extract was subjected to a differential Ammonium Sulfate (AS) precipitation. Dialyzed fractions containing the transposase were applied on a Heparin Sepharose FF (GE Healthcare) ion-exchange chromatography column and eluted in a NaCl gradient. M: molecular weight marker (Precision Plus Protein All Blue Standards (Bio-Rad))

4.2.1 Tnp binds to the β sliding clamp

To assay the interaction between Tnp and *E. coli* β, we covalently coupled the sliding clamp to tosyl-activated magnetic beads. Then, each transposase variant was incubated with the β-coated beads and immediately washed extensively to remove unbound Tnp. Reactions were stopped by adding SDS 1% and the eluted products analyzed by SDS-PAGE and revealed by Coomassie staining. Since SDS is used as reaction terminator and β is a dimer, any β monomers (β_1) not covalently linked to the magnetic beads are also eluted (Fig 4.13 B). To test the binding capacity of β coated beads we used GST-PolIV^{LF} and GST as positive and negative controls respectively, and showed that GST-PolIV^{LF} was retained by β coated beads. Conversely, as expected GST was not retained by β (data not shown), indicating that our system is suitable for the study of the potential interaction of Tnp with β.

When Tnp^{Wt} was assayed, results showed that up to 70% of Tnp^{Wt} used in the experiment is retained by β after extensive washes (Fig 4.13 B and C), concluding that both proteins effectively interact. We also aimed to map the binding site. There are many proteins in the cell that interact with sliding clamp and the interaction often depends on a small motif located at the N- or C-terminus of the protein. We have previously described that transposases belonging to different IS families interact with β through a binding motif preferentially located in the C-terminus of the protein (See 4.1.3). We therefore designed a C-terminal 20 amino acid truncation (named Tnp$^{C\Delta20}$) and tested its ability to bind β. Experiments showed that after several washes Tnp$^{C\Delta20}$ is still retained by β in a similar fashion than Tnp^{Wt} (Fig 4.13 B and C). We also purified a C-terminal 108 amino acid truncation (Tnp$^{C\Delta108}$) which deleted most of the C-terminal α-helix, and we showed that still bind β (data not shown). Hence, we dismissed the carboxyl-terminus of Tnp as the interacting region and focused our efforts in the study of the amino terminus of the protein. We purified the Inhibitor protein (Inh), a protein encoded by Tn*5* transposon with the same sequence as Tnp^{WT} but lacking 55 amino acids at the N-terminus of the enzyme. Our results showed that Inh is no longer able to bind β (Fig 4.13 B and C). To further circumscribe the binding region, we also purified an N-terminal 7 amino acid deletion (Tnp$^{N\Delta7}$) and demonstrated that it also failed to interact with β (Fig 4.13 B and C). We therefore conclude that the interaction with β resides at the extreme N-terminal 7 amino acids or, alternatively, that deletion of this region somehow alters the conformation of the protein in a way that is no longer able to interact with β.

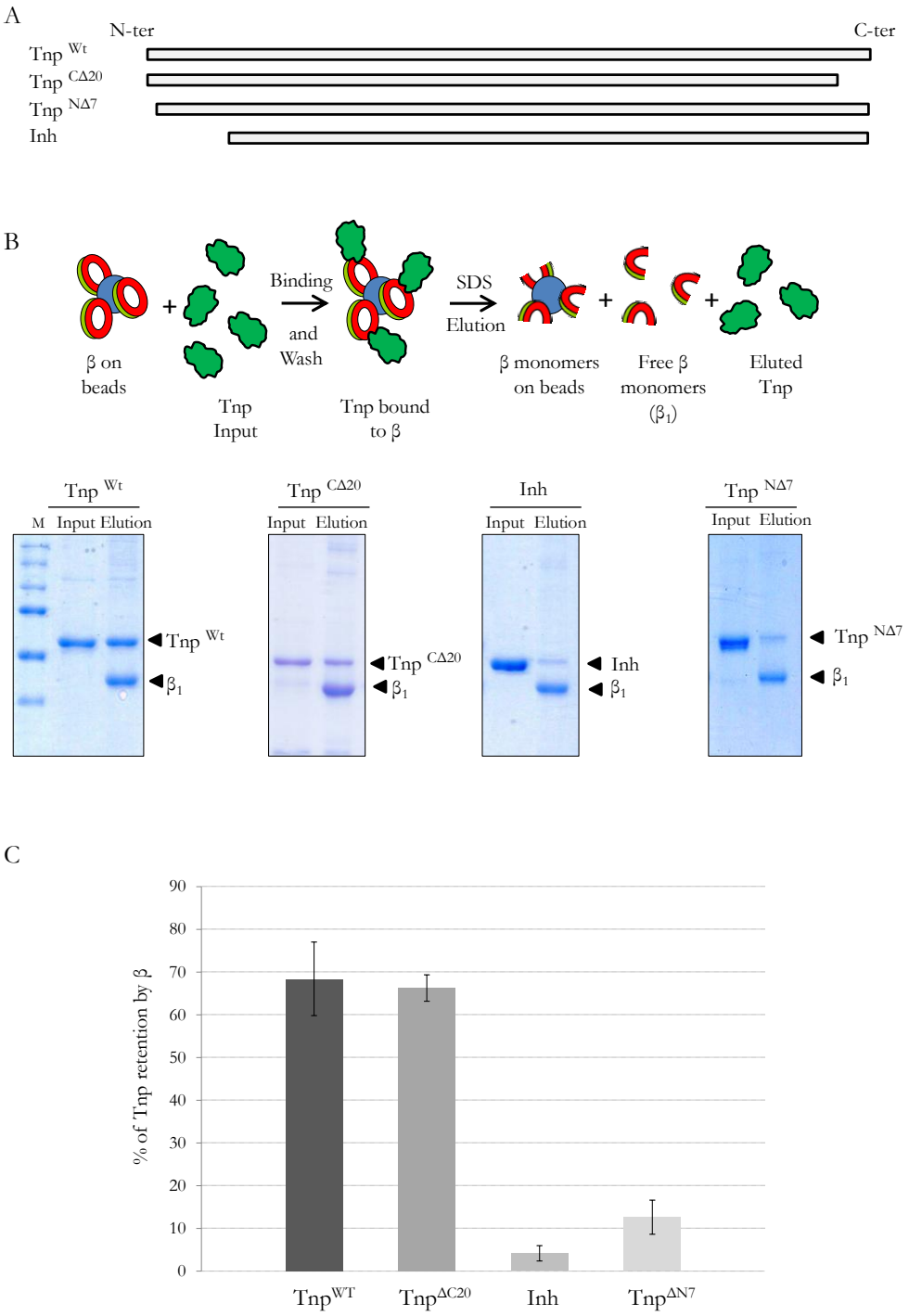


Figure 4.13. Tnp interacts with β sliding clamp.
A) Scheme of different Tnp truncations used in the binding assay. See Fig II.1 in Appendix II for amino acid sequences. B) At the top, diagram of the pull-down assay where β is covalently coupled to magnetic beads and proved for interaction with Tnp variants. Below, Coomassie stained SDS-PAGE gels of the interaction experiment. In each gel, first lane represent total amount of Tnp used in the assay (input) and second lane the eluted products (elution). From left to right, Tnp^{Wt} and $Tnp^{C\Delta20}$ are able to interact with β while Inh nor $Tnp^{N\Delta7}$ are not retained by β. M: molecular weight marker C) Densitometry of each Tnp input and eluted band are plotted as percentage of Tnp retained by β. Presented data are the result of three independent experiments, and standard deviations are also shown.

4.2.2 Tnp interaction with β promotes DNA binding

Next, we investigated whether the ability of Tnp^{WT} to bind DNA is affected by the presence of β. Purified Tnp^{WT} was incubated with a DNA 1kb PCR fragment containing 2 OE sequences (5´P- CTGTCTCTTATACACATCT-3´) either in the presence or absence of a fixed concentration of β. Reaction products were resolved by agarose gel electrophoresis, stained with SYBR Green and the DNA visualized with a UV trans-illuminator. Our results showed that Tnp^{Wt} alone has no significant effects on DNA electrophoretic mobility. However, when Tnp^{Wt} is incubated with DNA in the presence of β a new mobility retarded DNA complex is formed that is stable enough to be resolved in an electrophoresis gel (Fig 4.14 A and E). When $Tnp^{N\Delta7}$ is incubated with DNA, the complex is not formed in the presence or absence of β (Fig 4.14 B and E). These results suggest that β binding could induce a conformational change in Tnp that relieve the protein for its self-inhibitory state promoting DNA binding.

Transposition activity *in vitro* is measured as the rate of excised transposon or linear DNA donor backbone generated when a plasmid containing a Tn*5* transposon (or any DNA sequence flanked by OE) is incubated as a substrate of Tnp. Interestingly, Wt Tnp is inactive in these reactions and any activity *in vitro* has only been detected using hyperactive enzymes (Reznikoff 2008). We therefore explored if the addition of β had any effects on transposition reactions *in vitro* using Tnp^{Wt}. We used purified γ-complex to load β in a pSKT1-Tn*5* plasmid containing a target gene flanked by Tn*5* terminal repeats (see Fig. 4.15 A for plasmid design and terminal repeats sequences). β was loaded in the plasmid according to reaction conditions described elsewhere (Georgescu *et al.* 2008). Then, plasmid with loaded β was incubated with Tnp^{Wt} following the reaction conditions used in transposition assay with hyperactive mutants (Goryshin and Reznikoff 1998). However, after numerous attempts, we were not able to detected DNA transposition products under these conditions. Clearly, further investigation is required to elucidate why Tnp^{Wt} is inactive *in vitro*.

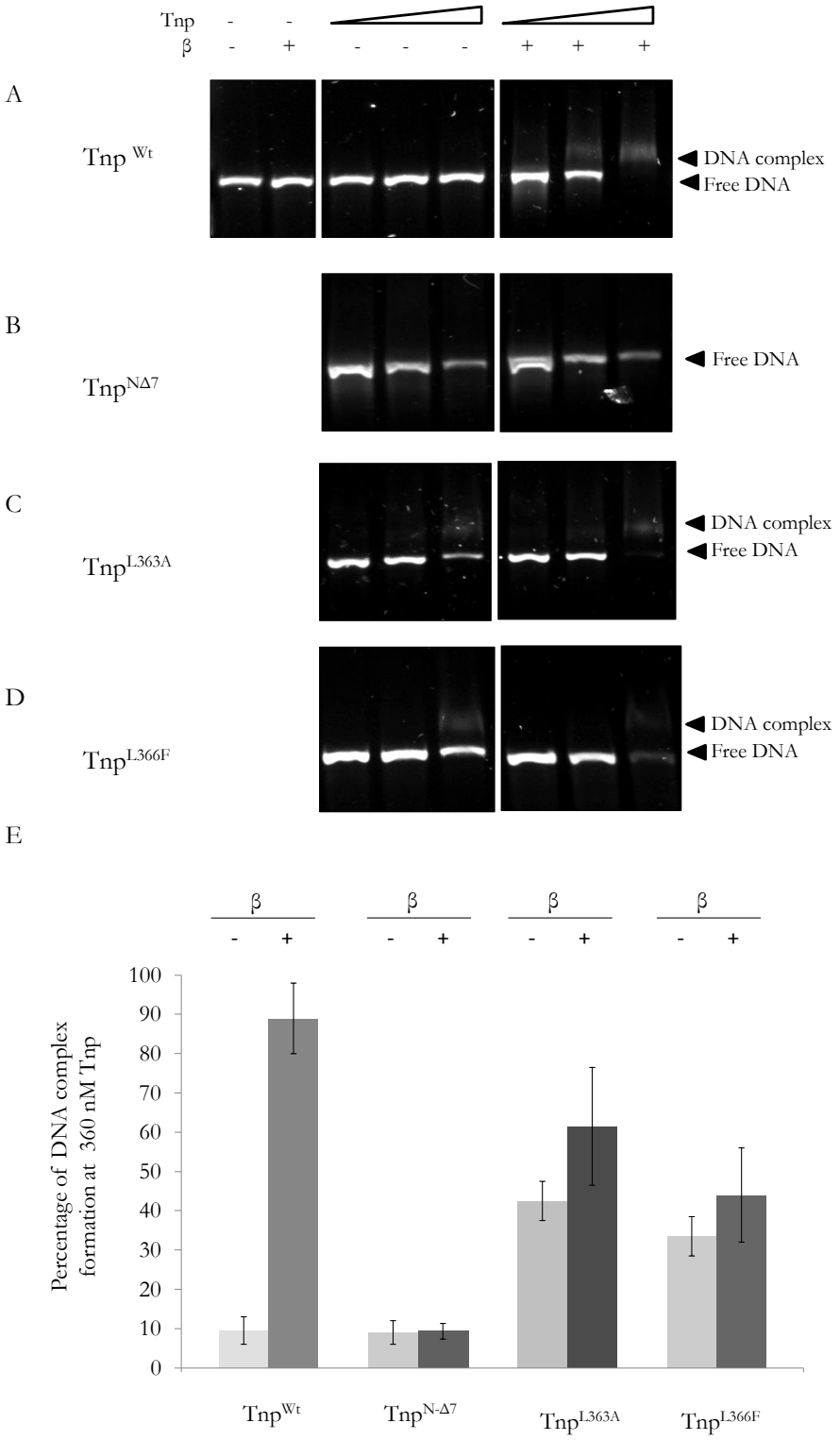


Figure 4.14. Tnp binds DNA in presence of β sliding clamp.
DNA mobilitiy shift assays. A tritation of Tnp^{Wt} and Tnp variants (90nM, 180 nM and 360 nM) are probed for interaction with DNA in the presence or absence of a fixed concentration of β. Tnp^{Wt} (A), $Tnp^{NΔ7}$ (B), Tnp^{L363A} (C) and Tnp^{L366F} (D) are assayed. Densitometry of free DNA and DNA complex bands at the 360 nM concentration of Tnp, are plotted as percentage of DNA complex formation (E).

4.2.3 Characterization of two novel hyperactive Tnp mutants

Because of its low transposition rate, the study of Tn*5* Tnp is linked to the discovery of hyperactive mutants. Described mutants only have a modest effect in increasing transposition activity *in vivo*. However, the combination of up to four different mutations raise the activity around 100-fold compared to Tnp^{Wt}. Here, we investigated residues in positions 363 and 366 that are highly conserved among Tn*5* transposase sequences of different organisms (Reznikoff *et al.* 2004). In order to study the effect in the transposase activity of point mutations L363A (Tnp^{L363A}) and L366F (Tnp^{L366F}), we performed an *in vivo* transposition assay. We used a recently-developed vector that generates genomic insertions of the lacZ gene flanked by the Tn*5* inverted repeats (Pajunen *et al.* 2010). The Tnp gene and its mutants were cloned under the transcriptional control of the BAD promoter, allowing for the modulation of its expression by addition of arabinose (Fig. 4.15 A). Transposase mediates the mobilization of the lacZ gene flanked by inverted repeats. When lacZ inserts into a gene of the bacterial chromosome in the correct orientation and reading frame, LacZ is transcribed and generates blue papillae. *E. coli* DH5α was transformed with plasmid constructions containing the Tnp^{Wt}, Tnp^{L363A} and Tnp^{L366F} genes. The vector with no cloned transposase was used as negative control. Cells were incubated 4 days at 37 ºC.
We observed papillae formation even in plates that did not contain arabinose, indicating that leakage in the expression of transposase was sufficient to generate transposition events. Surprisingly, a comparative analysis of the *in vivo* assay clearly reveals that Tnp^{L363A} and Tnp^{L366F} generate a much larger number of papillae than Tnp^{Wt}. Although blue spots are barely discernible particularly after three days of incubation in the two mutants, it could be estimated that papillae number are at least ~100-fold higher in both mutants with respect to Tnp^{Wt}. Moreover, Tnp^{L363A} and Tnp^{L366F} not only had an increased transposition activity, but also papillae appeared 24 h sooner than in Tnp^{Wt} (Fig. 4.15 B). The vector did not generate any papillae in our experimental settings, as expected. Hence, our data strongly support that Tnp^{L363A} and Tnp^{L366F} are hyperactive mutants, with a higher transposition activity *in vivo*.

Interestingly, hyperactive mutants here described, are able by themselves to interact with DNA and form the mobility retarded DNA complex even in the absence of β. Although Tnp^{L363A} and Tnp^{L366F} still bind to β (Fig. II.2 in Appendix II), the addition of the sliding clamp does not significantly improve DNA binding in these mutants (Fig 4.14 C and D). In light of these mutation locations, it could be hypothesized that L363A and L366F may induce a conformational change in the transposase that dislocate the contacts

between the N and C-terminal domains that are responsible of the self-imposed inhibition of the enzyme. Hence, it is tempting to speculate that Tnp^{Wt} requires β to be able to bind DNA, a limiting step in transposition, and in the cell Tnp will likely have to compete with other host proteins for β binding. In the other hand, Tnp^{L363A} or Tnp^{L366F} are β-independent to bind DNA which is reflected in 100-fold transposition rate increase in compare with Tnp^{Wt}.

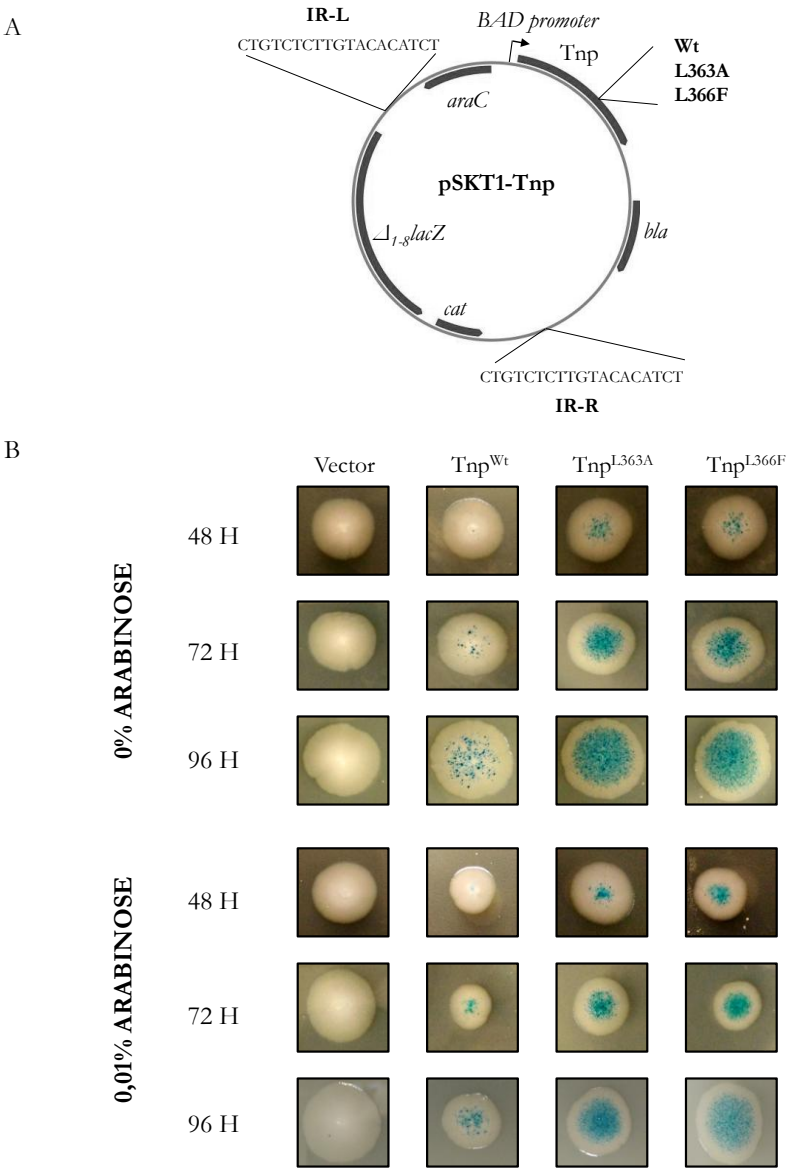


Figure 4.15. *In vivo* assay to study the transposition activity of different Tnp mutants
A) Plasmid design of vector pSKT1-Tn*5*. The sequences for the left (IR-L) or right (IR-R) inverted repeats are shown. We assayed the activity of Tnp^{Wt}, Tnp^{L363A} and Tnp^{L366F}. pSKT1 containing the inverted repeats but no transposase gene was used as a negative control. B) Papillation assay of *in vivo* transposition. The pictures shown representative examples of DH5α *E. coli* colonies of three different biological assays with four replicates each of them.

4.3 Transposase interaction with sliding clamp: effects on IS proliferation and transposition rate

Insertion sequences (IS) are ubiquitous in bacterial chromosomes and play a fundamental role in genome evolution. ISs have the ability to cross species barrier and transpose actively in new host, even far related. This characteristic makes ISs important players in horizontal gene transfer (HGT) processes. Additionally, IS dynamic models suggest that IS likely proliferate in chromosomes in abrupt burst rather than at constant rate (Olivier and Greene 2009). The invasion of a new host and IS expansion in the chromosome may rely in the compatibility of the transposase with its new molecular environment. We have previously described a widespread interaction of different IS families with the highly conserved replication factor sliding clamp. Here, we investigate to what extent this interaction limits o favor the ability of ISs to colonize a chromosome from a phylogenetically-distant organism and whether the strength of this interaction affects the transposition rate.

4.3.1 The transposase and related elements microarray

Although ISs are widely distributed among bacteria, genomes often contain significant amounts of truncated or inactive transposase sequences. In fact, few studies have discerned between those transposases that are active from those that remain inactive in the organism as a molecular fossil of past IS expansion and extinction cycles (Cerveau *et al.* 2011). Moreover, the analysis of mobile genetic elements dynamics entails difficulties and limitations by their sequence diversity, their presence in multiple copies in the chromosome and by their variable abundance within genomes, even between strains of the same organism. In order to study how interaction with sliding clamp affects transposition, we first aimed to develop a method that allows us to detect not only transposases that are actively transposing in a genome but also in natural communities.

Because of their atypical habitat conditions and low-biodiversity, acidic environments represent a perfect scenario for studying microbial evolution (López de Saro *et al.* 2013; López de Saro *et al.* 2015). Indeed, by proteomic techniques, transposon related proteins have been detected in acidic mine drainage (AMD) in biofilms formed primary by *Leptospirillum* (Ram *et al.* 2005). Besides, a *Leptospirillum* shotgun DNA microarray has also detected expression of some transposases in the acidic Tinto river (Parro *et al.* 2007). Hence, to gain insights in the dynamics of transposable elements in bacterial communities, we designed an oligonucleotide-based microarray consisting of probes representing 1,358

genes associated to ISs, phage and plasmid mobility functions present in the genomes of selected acidophilic organisms (Table 4.2). We chose prokaryotic organisms commonly found in Tinto river in Huelva (southwest Spain) because it is a well-characterized acidic ecosystem (Amaral-Zettler *et al.* 2011) what makes it an ideal benchmark for the analysis of mobile elements behaviour in bacterial populations.

Transposition is regulated by weak endogenous promoters (Nagy and Chandler 2004), thereby it is unlikely to detect high IS transcription levels in environmental samples. We used our developed microarray (see Materials and Methods for details) to detect expression of mobile elements in a natural ecosystem. We extracted total RNA form Tinto river water and we amplified it through a method based on a T7 RNA polymerase linear amplification (Moreno-Paz and Parro 2006). cDNA was synthesized form amplified anti sense RNA, labelled with Cy3-dUTP and hybridized against the microarray. We observed expression in most of the reference genes of the acidiophiles present both in the microarray and in the river. Although we detected a modest transcription of transposases in this sample, we were able to detect expression in four different IS families (IS*As1,* IS*200* ORF A, IS*200* ORF B and IS*21*) of *Leptospirillum ferrooxidans* RT32a, the most abundant bacteria in this environment, and also in a IS*21* element of *Acidiphilium sp. PM.* (Table 4.2). Thus, our microarray could be considered as a useful approach to study the expression of mobile elements in environmental communities.

Species	gyrB	rpoB	dnaX	Insertion Sequences
Acidiphilium cryptum JF5		+	+	
Acidiphilium sp. PM	+	+	+	IS*21*
Leptospirillum ferrooxidans LfeRT32a	+	+	+	IS*200* ORF A and B IS*As1* and IS*21*
L. rubarum		+	+	
L. group II 5way		+	+	IS*21*
L. ferrodiazotrophum	+	+	+	
Acidithiobacillus ferrooxidans 23270	+		+	
A. ferrooxidans 53993				
A. caldus		+		
Acidimicrobium ferrooxidans 10331	+		+	
Ferroplasma acidarmanus fer1	+	+		
Thermoplasma acidophilum	+	+	+	

Table 4.2. Expression of reference genes and IS-related genes detected in an acidic environmental sample with an oligonucleotide-based microarray
First column, organisms which transposases are represented in the microarray (Species). Reference genes (+) and IS families (as indicated) which transcription is detected in an environmental sample.

4.3.2 Detection of active IS in a long-term culture of *Acidiphilium* sp. PM

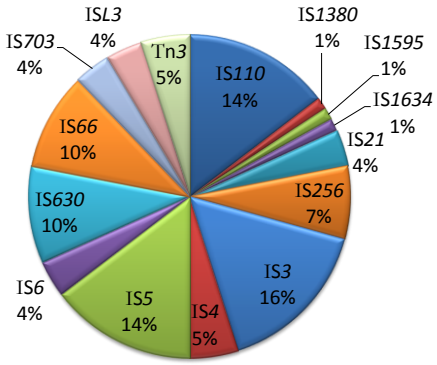
The sequencing of the genome of *Acidiphilium* sp. PM, an acidophilic α -proteobacterium isolated from the Tinto river, revealed that up to 2.4% of its chromosome is composed of repeated regions (San Martín-Uriz *et al.* 2011). We detected and classified by comparative analysis against the Pfam and ISfinder databases, the population of ISs in *Acidiphilium´*s genome, and found 85 non-redundant ORFs related to transposases and associated factors (Fig. 4.16 A). Although some of these ORFs corresponded to full-size transposases or associated proteins, many of them were likely inactivated remnants and gene fragments that are no longer capable of transposition. In order identify active transposases, we performed a long-term cultivation of *Acidiphilium* sp. PM. and analyzed changes in IS abundance that had become fixed in the population at the end of the experiment. The sequenced strain was used to start a culture which was grown for ~600 generations over four years and then we used our recently developed microarray, to identify changes in the IS population at the end of long-term culture. Although our methodological approach would not detect transposition events that did not change IS copy number in the chromosome (i.e. conservative transposition) or chromosomal rearrangements, we expected that it would detect IS proliferation or loss.

DNA obtained from the founder strain ('2007') and from the 4-year-old population ('2011') were differentially labelled with fluorescent dyes Cy3 and Cy5 and hybridized together against the microarray (Fig. 4.16 B). Results of the microarray represented in Fig. 4.16 C and Table III.1, are the average change of two different oligos representing each transposase and replicated both of them three times in each microarray. The microarray were assayed three independent times, one of them switching the fluorescent dyes Cy3 and Cy5 used to label the DNA from the two strains. We observed changes ($\log_2$['2011'/'2007']> 0.5) in nine elements (Fig. 4.16 C and Table III.1). Changes implied the apparent loss of the IS ORF in six cases (IS*4*-5*,* IS*110*-4*,* IS*110*-12*,* IS*21*-1*,* IS*3*-11, IS*5a*-2), and a small gain in copy number in three cases (IS*1634*-1*,* IS*5*b-8 and IS*66*-4).

To provide an overview of transcription at IS population level we used the microarray to detect expression of transposases in *Acidiphilium* sp. PM. Hence, we extracted total RNA of the culture, synthesized and fluorescently labelled second strand cDNA, so no amplification step was introduced. Interestingly when we hybridized cDNA from the starting point culture ("2007") against the microarray, the most significantly transcribed transposases we observed (IS*1634*, IS*21*-1, IS*5*a-2 and IS*66*-4) (Fig. 4.16 D) also have changed their copy number at the end of the long-term culture. When cDNA from "2007"

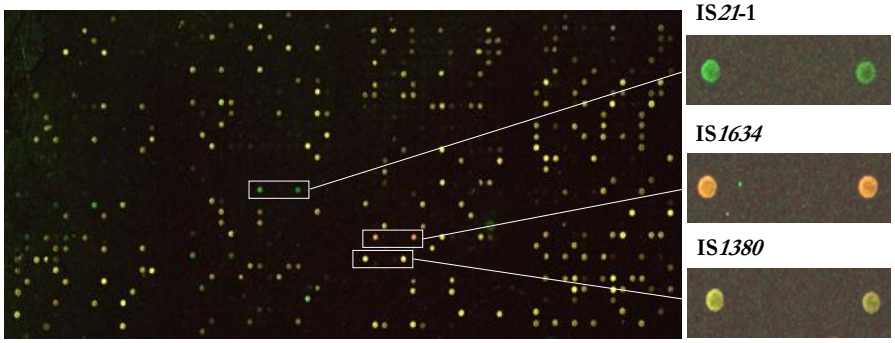
and "2011" cultures were differentially labelled and hybridized against the microarray, no significant differences in the expression pattern were observed, with the exception of those transposases that have disappeared in "2011" and consequently no transcription was detected.

A

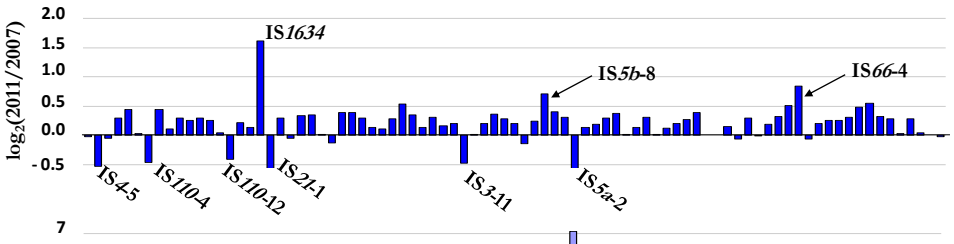


IS Family	Number of non-reduntant ORFS
IS*110*	12
IS*1380*	1
IS*1595*	1
IS*1634*	1
IS*21*	3
IS*256*	6
IS*3*	13
IS*4*	4
IS*5*	12
IS*6*	3
IS*630*	8
IS*66*	8
IS*703*	3
IS*L3*	3
Tn*3*	4

B



C



D

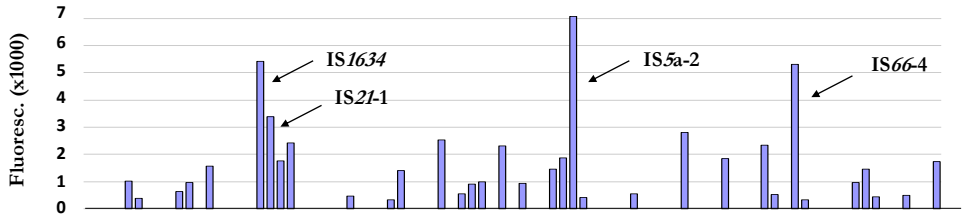


Figure 4.16. Identification of IS changes in a 600-generation culture of *Acidiphilium* sp.
A) IS families and number of non-redundant ORF present in the sequenced genome of *Acidiphilium* sp. PM. B) Oligonucleotide microarray to identify and quantify changes in ISs copy number in chromosomes. Three examples of are shown: for a IS that was lost (IS*21*-1), increased in copy number (IS*1634*-1), and remained without change (IS*1380*-1). C) Plot of the relative change

(log_2['2011'/'2007']) of transposases (or fragments) observed with the oligonucleotide microarray. Each bar represents the average change for each transposase, calculated from the fluorescence of two different oligos replicated three times in each microarray. The microarrays were assayed three independent times. Relevant ISs are indicated (see Table III.1 for the full list). D) Plot of the expression profile of IS population in the initial point of the long-term culture. Transposases significantly expressed are indicated.

4.3.3 An active transposase of the IS*1634* family contains a β-binding motif

Results obtained by the microarray, allowed us to indentify ISs that have proliferated or become extinguished in a long-term culture of a recently insolated microorganism. Although the modest global IS copy change detected, a member of the IS*1634* family (IS*1634*-1) present in a single copy in the '2007' culture, showed an increase to three copies after 600 generations. This revealed that IS*1634*-1 is an active transposase in the genome of *Acidiphilium,* therefore we focused on this mobile element for the rest of the study.

Because microarrays are semi-quantitative techniques, we also quantify the relative change of IS*1634* from "2007" to the "2011" culture by a quantitative PCR (qPCR) experiment. Known amounts of "2007" genomic DNA were serially diluted and used to create a standard curve for the IS*1634* transposase gene (r= 0.997). The genomic copy number of IS*1634* in "2011" time-point was determined from this "2007" standard curve, where the transposase is present in a single copy per chromosome. DnaX, encoding the γ subunit of DNA polymerase III, and present in a single copy in the chromosome was chosen as a reference. Results in Fig. 4.17 A, represent the average of "2011" dnaX and IS*1634* chromosomal gene copy number and the standard deviation (SD) of three replicates, and confirmed the previously observed increase by the microarray of IS*1634* in "2011".

Moreover, we performed an inverse PCR (iPCR) for further characterize the new insertions of IS*1634*. Genomic DNA was used to amplify the flanking regions of IS*1634* insertion sites both in "2007"and "2011" cultures. iPCR reaction products were resolved in a 1% agarose gel and also confirmed the IS*1634* copy number increase. In "2007" only one iPCR product was found but three in "2011" corresponding one of them to the original IS and the rest to the new IS insertions (Fig 4.17 B). These fragments were extracted from the gel and sequenced. Our analysis by sequencing of the insertion sites of the original and the new copies (named IS*1634*-2 and IS*1634*-3) revealed the characteristic direct repeats generated by transposition events, "AC**TA**GT" for IS*1634*-1, "TC**TA**AA" for IS*1634*-2 and "GA**TA**GA" for IS*1634*-3. We identified the consensus target site for this transposase as a "nC/A**TA**G/An" sequence.

However the study of the genomic context of the new insertions (Fig. 4.17 C) did not detect any obvious adaptive benefits for the evolving culture of *Acidiphilium* sp. In the other hand, IS*1634*-1 in the sequenced '2007' culture (Accession number: WP_007423974) could be a recent insertion in the chromosome, as suggested by its anomalous % GC and GC skew (Fig. 4.17 D).

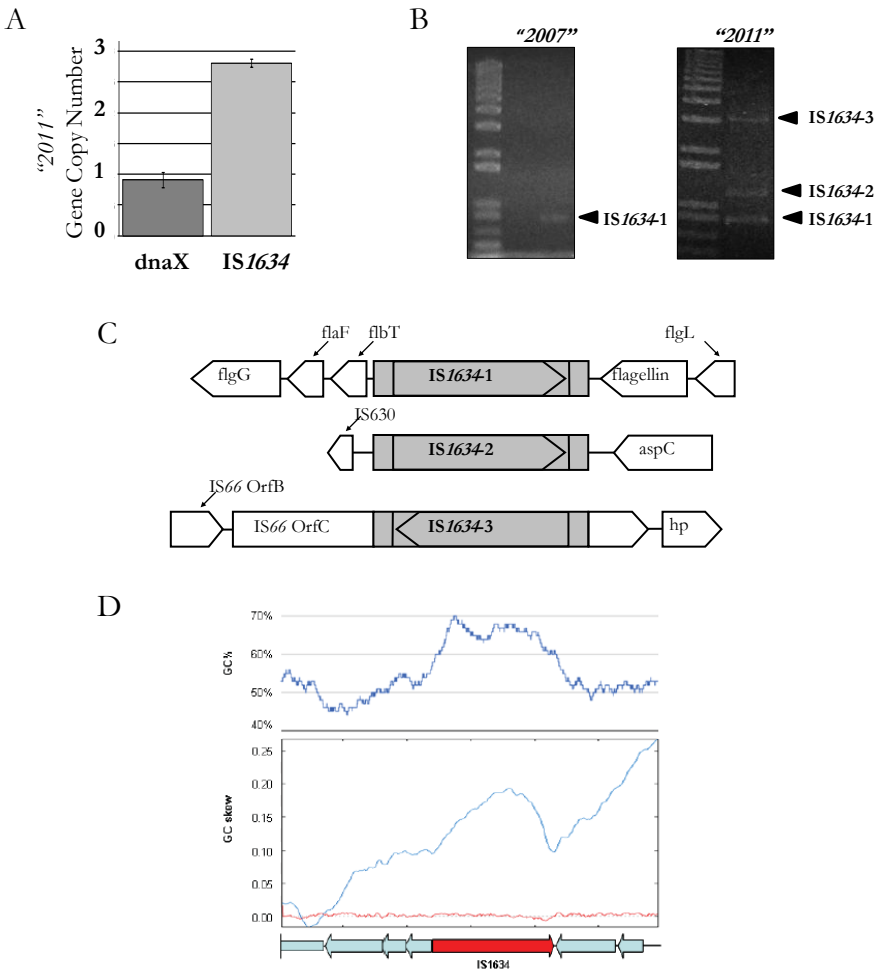


Figure 4.17. A member of the IS*1634* family is active in *Acidiphilium* sp. PM
A) qPCR experiment to quantify the relative change of IS*1634* from the '2007' to the '2011' culture. DnaX, encoding the γ subunit of DNA polymerase III, and present in single copy in the chromosome, was chosen as a reference. B) iPCR of IS*1634* transposase gen in "2007" and "2011" cultures using divergent oligonucleotides . iPCR products were resolved in a 1% agarose gel (arrow heads) C) iPCR sequencing showed the genomic context of IS*1634*-1 in the '2007' culture and the two additional copies present in the '2011' culture. IS*1634*-1 and IS*1634*-2 were located in the contig NZ_AFPR01000455 (10547 bp), while IS*1634*-3 was located in contig NZ_AFPR01000454 (3091 bp). 'hp', hypothetical protein. D) Genomic context of IS*1634*-1 in *Acidiphilium sp.* PM. Top panel, the content of G+C was calculated in the genomic region containing IS*1634*-1 using the program GC content calculator (www.biologicscorp.com) with a window of 500. Bottom panel, the cumulative GC skew (blue line), [(G -C)/(G + C)] was calculated using online GC skewing software (http://gcat.davidson.edu) with a window of 100 and a step size of 20. The region spans a region of 5986 bp.

The IS*1634* family is a poorly characterized family of insertion sequences related to IS*4*, IS*H3* and IS*701* (De Palmenaer *et al.* 2008). A similarity tree for the transposases encoded by IS*1634* members shows that this IS family can be found in Euryarchaea and in numerous groups of Bacteria (Fig. 4.18). Interestingly, IS*1634* appears in organisms that share acidic, heavy-metal rich environments, and is present, among others, in the chromosomes of most species of *Leptospirillum* (Nitrospira), *Acidithiobacillus* (γ -proteobacteria) and *Acidiphilium* (α -proteobacteria), as well as in *Paracoccus ferrooxidans* (α -proteobacteria), *Ferroplasma acidarmanus* (Euryarchaea), *Alicyclobacillus acidocaldarius* (Firmicutes), *Thiomonas sp.* (β-proteobacteria), and *Desulfobacca acetoxidans* (δ -proteobacteria) (Fig. 4.18). IS-sharing has been observed before for co-habiting organisms (Hooper *et al.* 2009).

The interaction of transposases with the host β sliding clamp occurs almost always at the C-terminus of the protein via a short sequence with a weak consensus sequence (Q_1-L_2-x-L_4-F_5). A search of the β-binding motif among IS*1634* transposases shows that a putative motif can be found at the C-terminus of these proteins (Fig. 4.18). Despite high variation in sequence context, the alignment of this region shows conservation of the Gln at position 1 of the putative β -binding motif plus conservation of hydrophobic or aromatic residues at positions 4 and 5, suggesting a possible interaction of IS*1634* transposases with the host β sliding clamp via this region.

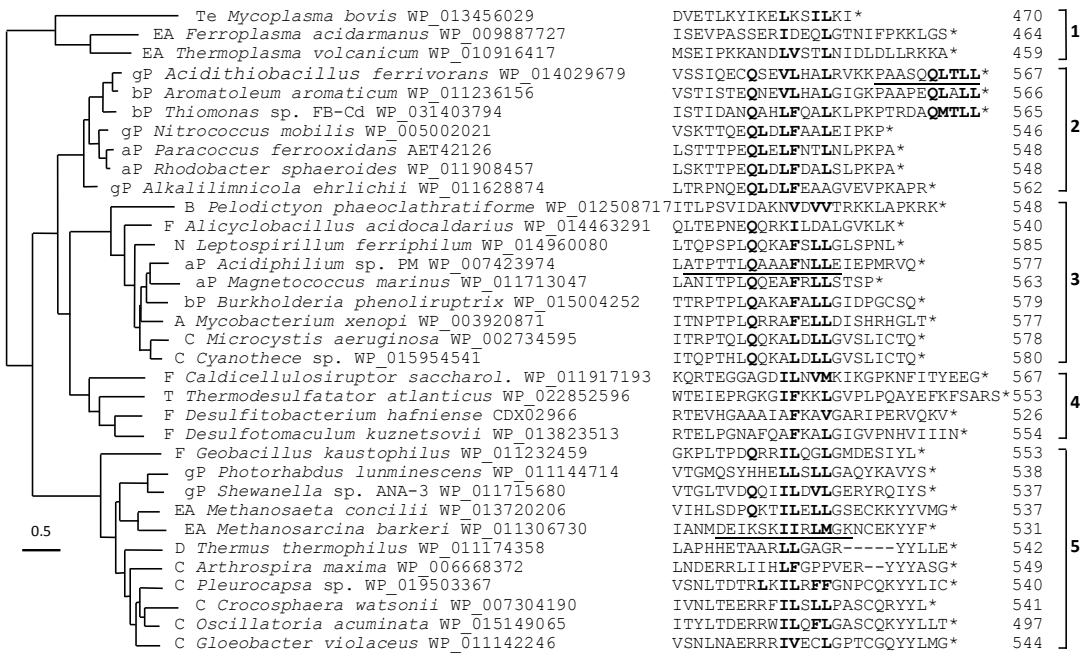


Figure 4.18. Unrooted similarity tree of IS*1634* transposases and alignment of C-terminal region.
IS*1634* transposase sequences were obtained following the methodology previously described, in which 2,216 genomes downloaded from the National Centre for Biotechnology Information (NCBI) Genome database on October 2012, were systematically scanned to identify and classify ISs. The collection of IS*1634* transposase sequences was made non-redundant at a sequence similarity level of 90% and by choosing one representative sequence per genus. The tree was computed with PhyML using a JTT model and a bootstrap of 500 replicates, and the results visualized with Seaview (Gouy *et al.* 2010). Bootstrap values are shown for the main branches. The analysis reveals five distinct groups of IS*1634* transposases. The code preceding each organism name is as follows: aP, bP, gP and dP stand for α , β , γ and δ Proteobacteria, respectively; F, Firmicutes; Te, Tenericutes; N, Nitrospira; C, Cyanobacteria; T, Thermodesulfobacteria; D, Deinococcus-Thermus; A, Actinobacteria; EA, Euryarchaeota. For each transposase the sequence of the C-terminus of the protein is presented and residues involved in the putative β -binding motif are aligned and in bold. Underlined sequences are those of peptides used for biochemical analysis in this work.

4.3.4 Transposases can bind β sliding clamps from diverse organisms

We have previously demonstrated using synthetic peptides that up to 9 transposases of different IS families of *E. coli* bind to β of the same organism. Besides, we established that the transposase of the *E. coli* Tn*5* transposon also binds to β. Here, we investigated not only the ability of the transposase of IS*1634* (Tnp) to bind to *Acidiphilium* β, but also we tested this interaction against other organism´s sliding clamps. This provide insights about whether this interaction could be maintained across species boundaries and to what extent this interaction limits or favors the ability of ISs to colonize a chromosome.

To determine whether *Acidiphilium* Tnp could bind β, we purified both proteins from overexpressing *E. coli* strains and probed their interaction biochemically. We purified the wild-type transposase and, for comparative purposes, two engineered mutant proteins. One in which the Q_1 and the F_4 of the β-binding motif had been mutated to Alanine (named Tnp^{5A}), disrupting the putative motif, and another one containing a full consensus sequence (QLSLF, termed Tnp^{CN}). These three transposases were purified from inclusion bodies and essentially followed the same purification procedure (See Methods and Fig. 4.19 for Tnp^{Wt} purification details).

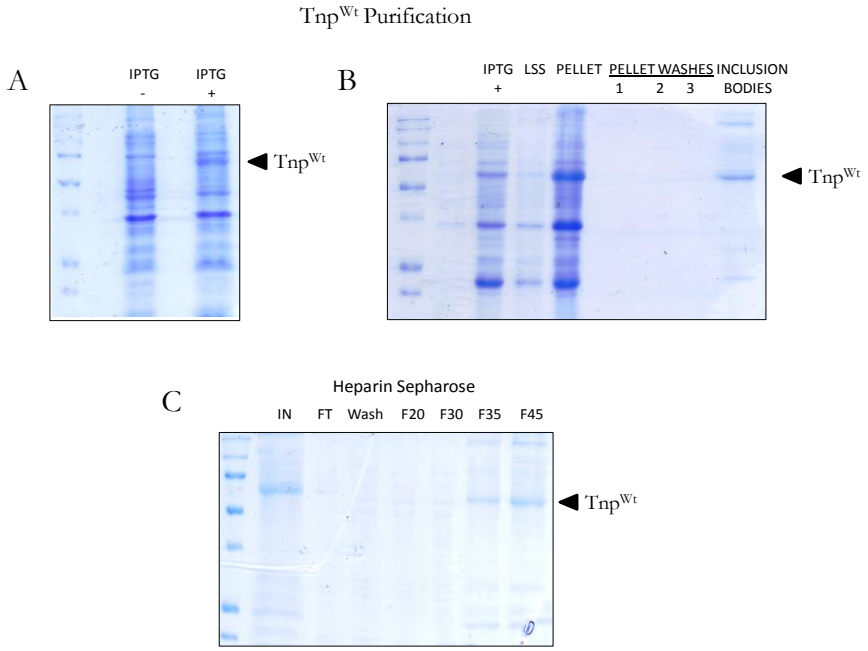


Figure 4.19. IS*1634* transposase (Tnp^{Wt}) purification.
A) Tnp^{Wt} was overexpressed in *E. coli* Rossetta® 2 (DE3) pLysS with 1 mM isopropyl β -D-1-thiogalactopyranoside (IPTG). First lane molecular weight marker, second lane un-induced control, third lane Tnp^{Wt} induced (IPTG +). B) The transposase overexpressed was insoluble and the pellet was washed three times with buffer supplemented with 1% Triton X-100. Inclusion bodies were dissolved in buffer containing 6M guanidine–HCl. This solution was diluted drop by drop in buffer without guanidine–HCl and the transposase was allowed to refold. Second lane induced Tnp^{Wt}, next protein extracts correspond to: soluble protein (Low speed supernatant, LSS), insoluble protein (cell debris, PELLET), three pellet washes with 1% Triton X-100 and last lane remaining Tnp^{Wt} in form of inclusion bodies after the washing protocol. C) Then it was applied on a Heparin Sepharose FF (GE Healthcare) ion-exchange chromatography column and eluted in a NaCl gradient. Lane 2 is the protein input (IN) in the column, lane 3 the flow-through (FT) and lane 4 an extensive wash of the column. Next lanes represent fractions of the NaCl gradient as indicated.

In addition to *Acidiphilium* β protein (Acβ), we purified β from *Leptospirillum ferrooxidans* (Lfβ) (Fig 4.20) and *Escherichia coli* (Ecβ) (Fig 4.5). We chose *Leptospirillum* (phylum Nitrospira) because, although distant from *Acidiphilium* phylogenetically, it shares the same acidic habitats, where it is usually the most common organism (Amaral-Zettler *et al.* 2011). This represents a suitable scenario to study the role that could play the transposase-β interaction in the horizontal IS transfer between bacteria sharing the same ecosystem. *E. coli* was chosen because its β sliding clamp and its ligands have been well described, facilitating the performance of comparative and genetic experiments as described below, and because no IS*1634* member has been yet detected in Enterobacteria.

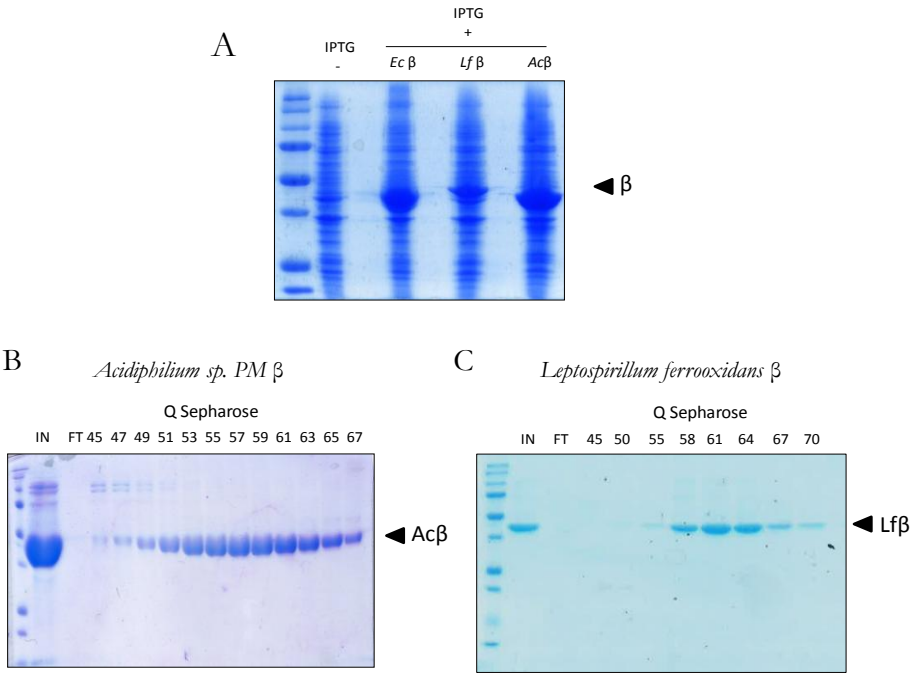


Figure 4.20. *Acidiphilium sp.* and *Leptospirillum ferrooxidans* β sliding clamps purification.
A) β of *E. coli*, *Leptospirillum ferooxidans* and *Acidiphilium sp.* were overexpressed (lanes 3, 4 and 5 respectively) in *E.coli* BL21 with 1mM isopropyl β -D-1-thiogalactopyranoside (IPTG). Lane 2 represents a protein extract before induction. *Leptospirillum* and *Acidiphilium* β formed inclusion bodies and they were treated as described in Methods. *E. coli* β purification is detailed in Fig 4.5 B) Refolded *Acidiphilium* β was applied on a Q Sepharose (GE Healthcare) ion-exchange column (IN) and fractions from a NaCl gradient collected as indicated. C) β of *Leptospirillum* was applied on a Q Sepharose column and proceeded as in B.

We assayed the interaction by covalently coupling Tnp to tosyl-activated magnetic beads and probing them with fluorescently-labelled β. The fluorescent label of Acβ and Lfβ followed the same principles of *E. coli* β labelling previously described. After extensive washing to remove unbound β, the reaction was stopped with SDS 1%, and loaded in a SDS-PAGE gel. The retention of β by the transposases was analyzed on a UV transilluminator. As shown in Fig. 4.21 A, Tnp^{Wt} can bind and retain β from *Acidiphilium*, *Leptospirillum* and *Escherichia* but the double mutant (Tnp^{5A}) could not. The consensus mutant, Tnp^{CN}, could bind to β from the three species.

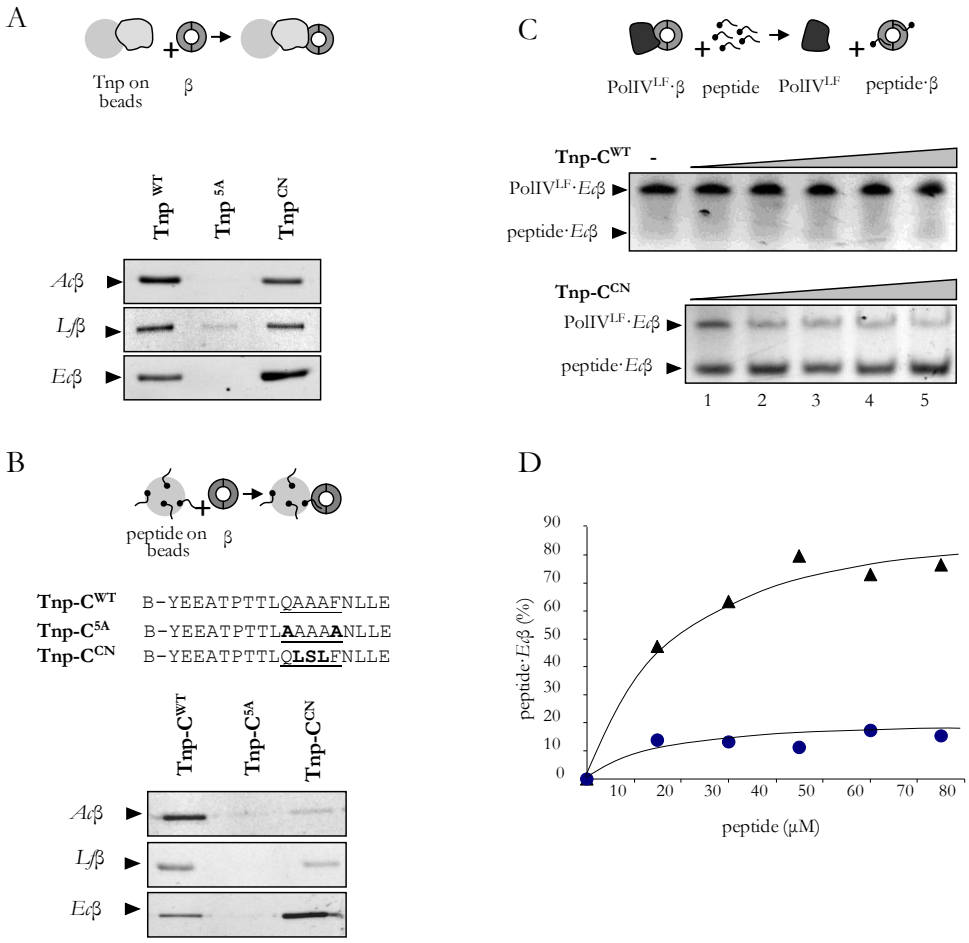


Figure 4.21. Interaction between Tnp and sliding clamps.
A) Binding assay of *Acidiphilium* Tnp^{Wt} and mutants Tnp^{5A} and Tnp^{CN} to labelled β. From, top to bottom, *Acidiphilium, Leptospirillum* and *Escherichia coli* β. Tnps was covalently coupled to tosyl-activated magnetic beads and any β retained analyzed by SDS-PAGE. B) Binding assay of peptides derived from C-terminal region of Tnp^{Wt}, Tnp^{5A} and Tnp^{CN} of *Acidiphilium*. At the up of the panel peptide amino acid sequences are showed and putative β -binding motif are underlined. Those amino acids changed from the Tnp^{Wt} sequence are in bold type; B, biotin. The N-terminally biotinylated peptides were bound to streptavidin beads and assayed with labelled *Ac*β, *Lf*β and *Ec*β. Retained β was analyzed by SDS-PAGE. C) Competition assay of complexes of *E. coli* β bound to the C-terminal domain ('little finger', LF) of DNA polymerase IV (PolIV^{LF}). The Tnp-C^{Wt} (top panel) and Tnp-C^{CN} (bottom panel) peptides derived from *Acidiphilium* were used to disrupt PolIV^{LF}·Ecβ complexes. Concentration of peptides is as follows for lanes 1–5: 15, 30, 45, 60, 75 μM. D) Quantification of PolIV^{LF}·Ecβ and peptide·Ecβ complexes of (C) was done by densitometry and plotted as percentage of peptide·Ecβ bound. Tnp-C^{WT} data is represented with circles and Tnp-C^{CN} is represented as triangles.

In order to further demonstrate that the C-terminal regions of Tnp contained the β-binding motif, we synthesized peptides containing these sequences. The peptides, which contained biotin at the N-terminus, were coupled to streptavidin-coated magnetic beads and probed against fluorescently-labelled β from the three species. Results in Fig. 4.21 B

show that the peptide containing the wild type β -binding motif (named Tnp-C^{Wt}) can retain β from the three species. Similarly, the peptide with the consensus motif (Tnp-C^{CN}) could bind to all β. However the double mutant peptide (Tnp-C^{5A}) cannot bind to any sliding clamp. The data from these two experiments confirm, first, that Tnp of *Acidiphilium* interacts not only with *Acidiphilium* β, but also with the sliding clamp of other two different species; and second, that the C-terminus of the transposase is responsible for the interaction with β by a putative β binding motif previously described.

Since experiments in Fig. 21 A and B are mostly qualitative, we used a competition assay to analyze the relative strength of the interaction between the Tnp-C^{Wt} and Tnp-C^{CN} peptides and Ecβ. This assay makes use of the gel-shift generated by DNA polymerase IV (LF domain, $PolIV^{LF}$) bound to β in native gel electrophoresis. These two proteins bind strongly and their interaction has been described in atomic detail. $PolIV^{LF}$ binds to Ecβ in the canonical hydrophobic pocket on the C-side of the ring which is the binding site of all other proteins studied to date (Bunting *et al.* 2003). Therefore a displacement of this complex by an excess of peptide, observable in the native gel by a change in the mobility of β, would indicate that the peptide binds strongly to the same pocket on β. A titration of the Tnp-C^{Wt} and Tnp-C^{CN} peptides on the preformed complex Ecβ:$PolIV^{LF}$ and a quantification of the shift in the β mobility, indicate that the Tnp-C^{CN} peptide binds to Ecβ with higher affinity than the Tnp-C^{Wt} peptide (Fig. 4.21 C and D).

In addition, to further characterize the biochemical interaction between Tnp and Ecβ, we performed a chemical crosslinking of Tnp and β. Tnp in the presence of crosslinker adopted a tetrameric conformation in accordance with its molecular weight in a SDS-PAGE gel (Fig. 4.22). This conformation has been postulated before for other transposases (Dyda *et al.* 2012). The crosslinking of Tnp and Ecβ revealed a complex that is consistent with a stoichiometry of a tetramer of Tnp interacting with two β dimmers.

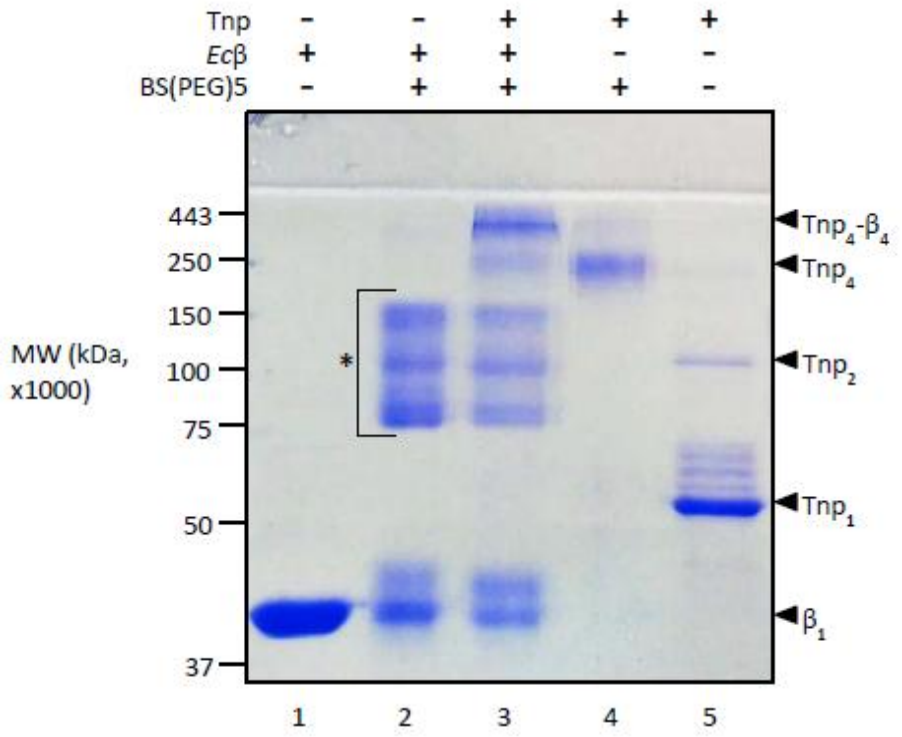


Figure 4.22. Chemical crosslinking of *Acidiphilium* Tnp^{Wt} and Ecβ.
SDS-PAGE of Ecβ and *Acidiphilium* Tnp^{Wt}. In lanes 1 and 5, β and Tnp, respectively, migrate according to their expected molecular weights in the absence of the crosslinker. BS(PEG)5 crosslinking reveals that β readily crosslinks in dimers, trimers and tetramers (lane 2, shown with an asterisk). In the presence of BS(PEG)5, *Acidiphilium* Tnp crosslinks into a form that is consistent with a tetrameric transposase (lane 4). When β and Tnp are mixed in the presence of crosslinker (lane3) there is a new complex formed which is consistent with tetrameric Tnp and two dimmers of β.

The alignment of Fig. 4.18 showed that in some species (*Aromatoleum*, *Thiomonas*, *Desulfobacca* and *Acidithiobacillus*) a potential second motif can be found at the C-terminal end of some *IS1634* family transposases. To determine whether these sequences could bind to β, we synthesized a peptide containing the C-terminal sequence of *Acidithiobacillus ferrivorans* IS*1634* transposase. As shown in Fig. 4.23 A, this peptide (*Af*Tnp-C^{Wt}) can bind and retain *Acidiphilium*, *Leptospirillum* and *E. coli* β, while a double mutant in which the Q and L in first and fourth positions of the motif were mutated to alanine (*Af*Tnp-C^{Mt}), cannot. Further, *Af*Tnp-C^{Wt} can displace PolIV^{LF} from a complex with *Ec*β , revealing that the *Acidithiobacillus* peptide binds to the same region of *Ec*β as other peptides (Fig. 4.23 B).

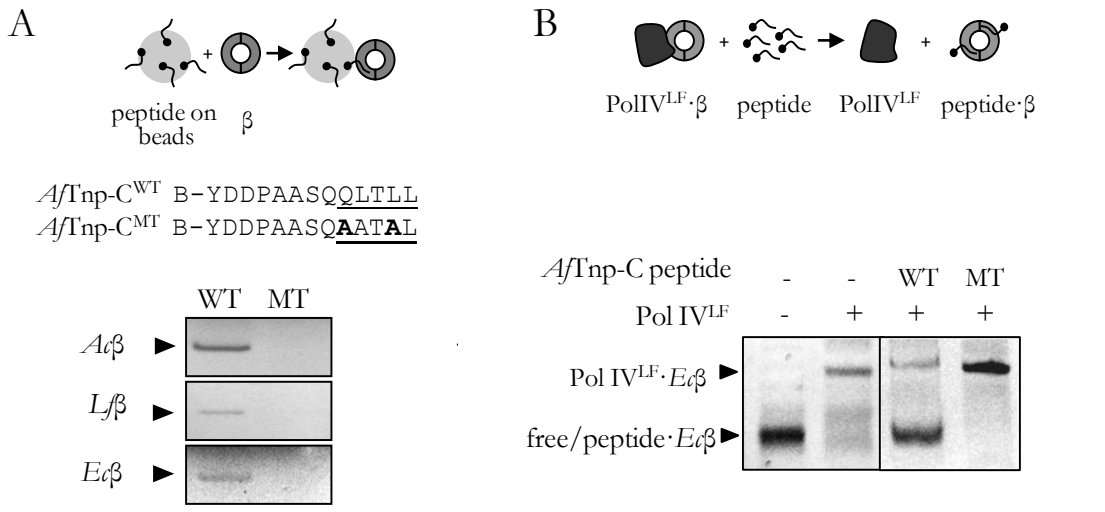


Figure 4.23. Binding of *Acidithiobacillus* IS*1634* transposase to sliding clamps.
A) Interaction of peptides derived from *Acidithiobacillus ferrivorans* IS*1634* transposase (WP_014029679) (*Af*Tnp-C) with β from *Acidiphilium*, *Leptospirillum* and *E. coli*. Peptide sequences are shown for the wild-type sequence (*Af*Tnp-C^{Wt}) and mutant (*Af*Tnp-C^{Mt}) and the putative β -binding motif underlined. The sequence of the C-terminus of *Acidithiobacillus* transposase starts at residue 4 of the synthetic peptide (see Fig. 4.18). The biotinylated (B) peptides were bound to streptavidin-coated magnetic beads and assayed with the different β. B) Native PAGE was used to resolve PolIV^{LF}·β complexes from free or peptide-bound β, demonstrating that *Af*Tnp-C^{Wt} peptide binds to the same hydrophobic pocket on β as PolIV^{LF}.

To provide more evidences that the ability to interact with sliding clamps of different organism is a general mechanism applicable to different IS families, we looked for more interspecies Tnp-β interactions. We detected that a member of the IS*66* family (IS*66*-4), also increased its copy number during the 600 generations long-term culture of *Acidiphilium* (Fig. 4.16 C), indicating that it is active. A sequence alignment of transposases of the IS*66* family, including *Acidiphilium* IS*66*-4 and *E. coli* IS*66* TnpB (YP_424826 and Fig. 4.4) revealed a putative β binding motif in the C-terminal of the protein (Fig 4.24 A). We already probed that *E. coli* IS*66* TnpB interacts with *E. coli* β (Fig. 4.6 and Fig. 4.7), and now we tested this interaction with *Acidiphilium* and *Leptospirillum* β. As is shown in figure 4.24 B, a peptide derived from *E. coli* IS*66* TnpB retained β of both acidiphilic bacteria.

A

```
Escherichia coli (YP_424826) IS66 TnpB   79 PVTR--DGKVHLTPAQLSMLLEGINWK--HPKRTERAGIRI*
Acidiphilium sp (WP_007421356)IS66-4    181 PQSG--TTLMSLSPAQLATLLEGCEWR--APVQSLRPVLAG*
Acidiphilium sp (WP_007421581)           79 PQTA--DGVVFLTAGQIGYLLEGIDWR--NPQQTWRPQAAG*
A. ferrooxidans (YP_002218731)           78 PRAD--AGALELSAAQWAMLVEGRPWTPLPTLEKCTPKLL*
Mesorhizobium ciceri (WP_027035507)      79 PVMAGFDGSITLTPAQLAMLIEGIDWR--APERVWRPALAG*
Methylocapsa acidiphila (WP_026607565)   79 PRMAGFEGSITLSPAQLAMLIEGIDWR--IPERVWRPAIAG*
Azospirillum brasilense (WP_035681092)   79 PISR--EGVAVLTPAQLAMLLEGMDWR--APQRPGRPEMAG*
Sphingomonas sp. (EGI53119)              74 PVTA--TGTVTLTPAQLSMLLEGIDWR--RPERTFTPTLAG*
Bradyrhizobium sp.(WP_008545089)         79 PSSA--DGVVTITPAQLGYLLEGIDWR--MPQQTWRPQAAG*
M. magneticum(WP_011382849)              77 PSPA--DGIVGLTPAQLGMLLEGIDWR--MPIRTWKPQSAG*
Pseudomonas fluorescens (WP_020723502)   79 PQAT--SGSVSLTAAQLSMLLEGIDWR--RPIRT-APVLAV*
Variovorax paradoxus (WP_012745822)      79 PQAT--SGSVSLTPAQLSMLLEGIDWR--MPVRTHEPLLAA*
```

B

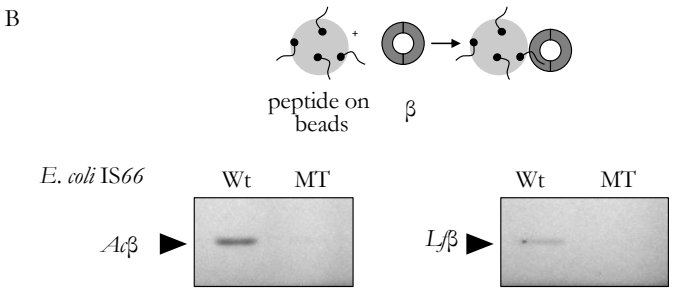


Figure 4.24. *E. coli* IS*66* transposase also binds sliding clamps from diverse organisms
A) Alignment of C-terminal sequences of IS*66* transposases, including *Acidiphilium* IS*66*-4. A *E. coli* peptide used in biochemical analysis is boxed, and residues putatively corresponding to the β binding motif are in bold. B) A biotinylated *E. coli* IS*66* TnpB peptide bound to streptavidin-coated magnetic beads is able to retain *Ac*β and *Lf*β.

4.3.5 Transposases can bind bidirectionally to β and to the Archaeal PCNA

We have also detected a sliding clamp-binding motif in transposases of two IS*1634* members from Euryarchaea (*Methanosarcina* and *Methanosaeta*) (Fig. 4.18) compatible with the consensus PCNA-binding motif (Q-x-x-I-x-x-F-F) (Warbrick 2000). PCNA (Proliferating Cell Nuclear Antigen), the Archaeal and Eukaryotic sliding clamp, is a trimmer in solution and functionally homologous to bacterial β. We demonstrated that a biotinylated peptide derived from the IS*1634* transposase present in *Methanosarcina barkeri* (*Mb*Tnp-C) and containing the putative PCNA-binding motif, retained purified and labeled *M. barkeri* PCNA (Fig. 4.25 A).

Moreover, given the relatedness and overlap between the β and PCNA binding motifs, we also tested whether these interactions could be maintained across phylogenetic domains. We observed that the peptide derived from the *Methanosarcina* transposase, *Mb*Tnp-C, also retained *Ac*β (Fig. 4.25 A). Then, we tested whether the *Acidiphilium* Tnp^{Wt}

could interact with *Methanosarcina* PCNA. As shown in Fig. 4.25 B, Tnp^{Wt} retained PCNA, while the mutant Tnp^{5A} did not. These results shown that the interaction transposase-sliding clamp could potentially be maintained, bidirectionally, across the Bacteria-Archaea boundary.

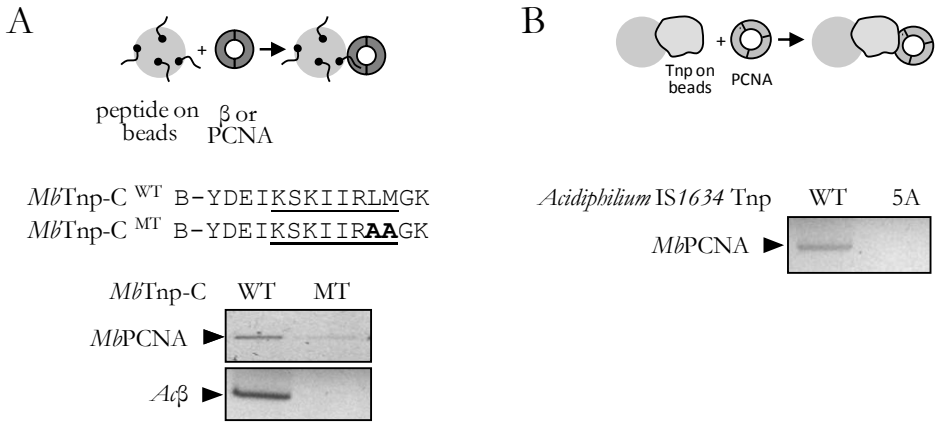


Figure 4.25. **Binding of transposases to Archaeal sliding clamp (PCNA)**
A) Interaction of peptides derived from the *Methanosarcina barkeri* IS*1634* transposase (WP_011306730) (*Mb*Tnp-C) with *Mb*PCNA (upper panel) and *Ac*β. (lower panel) B) Interaction between *Mb*PCNA and *Acidiphilium* IS*1634* Tnp^{Wt} and Tnp^{5A}. Tnp^{Wt} coupled to tosylactivated magnetic beads is able to retain labelled *Mb*PCNA, however Tnp^{5A} is not.

4.3.6 A stronger β-binding motif increases transposition in vivo

Although the β-binding motif is universal, the established consensus sequence (QLxLF) was experimentally validated using *Escherichia coli* β (Dalrymple *et al.* 2003; Georgescu *et al.* 2008b). A comparison of several β-binding motifs found in *E. coli* and *Acidiphilium* DNA replication and DNA repair enzymes shows that those of *E. coli* better conform to the consensus (Fig. 4.26). Besides, the identity between *Acidiphilium* and *E. coli* β is only 35.4%. These indicate the existence of subtle differences in the motif of β-binding proteins between both organisms in order to better bind their own β. Although *Acidiphilium* Tnp^{Wt} binds *E. coli* β (Fig. 4.21 and Fig 4.22), we have also demonstrated that a Tnp variant containing the β -binding consensus motif (Tnp^{Cn}) binds *E. coli* β with higher affinity than the Tnp^{Wt} (Fig. 4.21 C, D). This arise the question whether this higher affinity would result in more efficient transposition.

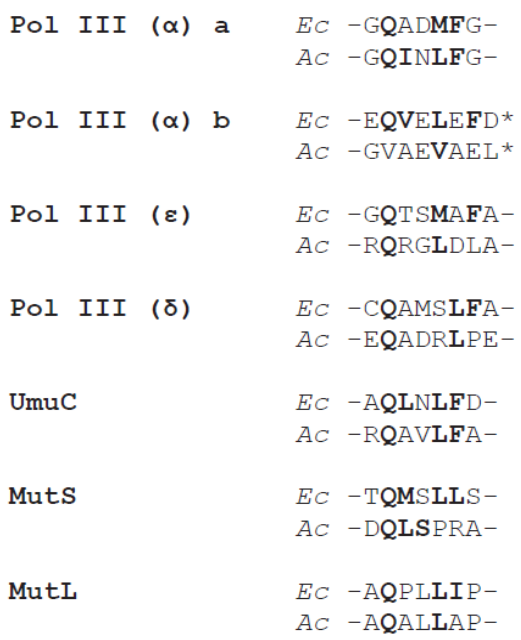


Figure 4.26. β-binding motifs present in enzymes of *E. coli* and *Acidiphilium sp.* PM. Residues shown correspond to the β-binding motifs present on the α (a: internal; b: C-terminal), ε and δ subunits of DNA polymerase III; UmuC (DNA polymerase V); and the mismatch repair enzymes MutS and MutL. These enzymes were fully aligned but only the section corresponding to the β-binding motifs is shown. *E. coli* (*Ec*) and *Acidiphilium* (*Ac*).

In order to determine whether the interaction detected *in vitro* with β has an effect in the ability of transposases to be functional in other organisms, we performed an *in vivo* transposition assay for the three variants of the *Acidiphilium* Tnp (Tnp^{Wt}, Tnp^{Cn} and Tnp^{5A}) in *E. coli*. We used a vector that generates genomic insertions of the lacZ gene flanked by the IS inverted repeats (Pajunen *et al.* 2010). The *Acidiphilium* IS*1634* transposase gene was placed under the transcriptional control of the BAD promoter, allowing for the modulation of its expression by addition of arabinose (Fig. 4.27 A). We transformed *E. coli DH5a* cells with the three plasmid variants containing Tnp^{Wt}, Tnp^{Cn} and Tnp^{5A} genes, plus a control plasmid with no transposase cloned, and incubated for 15 days at 30 °C. We observed papillae reflecting transposition events in Tnp^{Wt} colonies, although clearly occurring at a low frequency in our experimental conditions if compared for example, with papillae number of Tn*5* (Fig. 4.15). Low papillae number could be explained because IS*1634* gene has abundant rare codons for *E.coli* (Fig. III.2 in Appendix III), which could be reflected in low transcriptions levels. Mutant Tnp^{5A} showed lower generation of papillae than Tnp^{Wt} but further experimentation will be required to investigate whether this mutation decreases or eliminates transposition. However, cells harbouring the plasmid with the consensus mutation, Tnp^{Cn}, clearly had a significant increase of transposition with respect to Tnp^{Wt},

generating ~4–6-fold more papillae (Fig. 4.27 B and C). Our results therefore suggest that stronger binding of the transposase to β increases the frequency of transposition events in the cell.

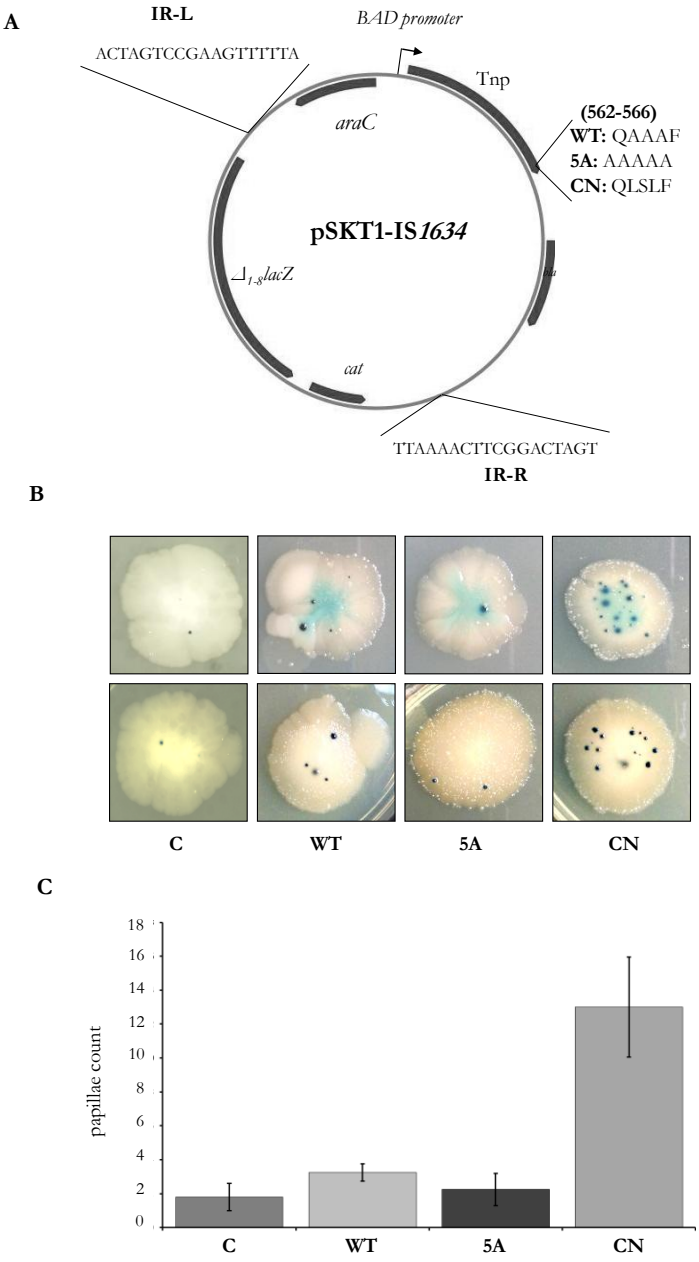


Figure 4.27. *In vivo* transposition assay to study the effect of different β binding motifs in IS*1634* transposition rate.

A) Plasmid design of vector pSKT1-IS*1634* (See Methods). The sequences for the left (IR-L) or right (IR-R) inverted repeats is shown. The sequences of the wild type and mutants 5A and CN is shown (amino acids 562–566 of IS*1634* Tnp). B) Papillation assay of transposition. The pictures show representative examples of *E. coli* colonies for the three versions of the transposase (WT and 5A or CN mutants), and a negative control C (pSKT1 containing the inverted repeats but no transposase gene). C) Quantification of the papillation assay for 8 colonies (C = 1.51 ± 0.6; WT = 3.15 ± 0.5; 5A = 2.05 ± 0.85; CN = 13.0 ± 2,95).

4.4. PCNA-binding motif in the human retrotransposon LINE-1

We have already shown that interaction between transposase and sliding clamp is a Phylogenetically wide distributed mechanism linking replication and transposition. We have studied this connection mainly between DNA transposons (class II transposable elements) and bacterial β sliding clamp, but we have also provided data that extends this interaction to Archaean PCNA. To draw a broader picture of the transposition-replication relationship, here we investigate how a retrotransposon (class I of transposable elements) interacts with a Eukaryotic sliding clamp.

The sequencing of the human genome revealed that up of the 45% of our genome is comprised by genetic mobile elements or derivates (Landers *et al.* 2001, de Koning *et al.* 2011). DNA transposons in humans account for the 3% of the genome and are currently inactive, remaining in the genome as molecular fossils; however it has been proposed that they were highly active during primate linage evolution (Pace *et al.* 2007). In the other hand, within retrotransposons there are two groups, classified by the presence or not of Long Terminal Repeats (LTR), being the non-LTR group the most abundant in the genome, accounting by one-third of the human genome. Non-LTR retrotransposons includes Long Interspersed Element-1 (LINE-1) and Short Interspersed Elements (SINEs such as Alu, and SVA elements). LINE-1 represents the 17% of the genome and are active autonomous elements that transpose via a RNA intermediate (Sassaman *et al.* 1997). They are also able to mobilize non-autonomous elements like SINEs. In the human genome there is an estimated of 500000 LINE -1 copies from which only around 100 of them are competent for transposition (Brouha *et al.* 2003).

An active human LINE-1 is composed by two open reading frames (ORF 1 and 2). Both ORFs are flanked by a 5´untranslated region (UTR) and a 3´UTR. In the 5´UTR is located a promoter from where ORFs are transcribed, and another antisense promoter (Speek *et al.* 2001). The 3´UTR ends in a tail of variable length rich in oligo (dA). The ORF1 encodes a RNA binding protein with chaperone activity (Martin and Bushman 2001, Babushok *et al.* 2007) that bind RNA in form of a homotrimeric protein (Khazina *et al.* 2011). Meanwhile, the ORF2 encodes a protein with endonuclease and reverse-transcriptase activities (Babushok *et al.* 2007). Between these mayor domains, there is another one with no clear function named Z segment (Clements and Singer 1998). Besides ORF2 protein has a Cys rich region in the Carboxi- termini of the protein which function, although unclear, seems to be implicated in RNA binding (Piskareva *et al.* 2013) (Fig 4.28)

LINE-1 transpose in a "copy-paste" mechanism that is not fully understood. Briefly, LINE-1 is transcribed from its own promoter by host RNA polymerase II and the bicistronic mRNA is translated in the cytoplasm in ORF1 and ORF2 proteins (Lavie *et al.* 2004), both required for effective transposition (Feng *et al.* 1996). ORF1 and ORF2 bind RNA preferentially in cis (i.e. the mRNA from they have been translated) (Wei *et al.* 2001) and the resulting ribonucleoprotein complex is transported to the nucleus. Here, in a mechanism named *target-site primed reverse transcription*, ORF2 with its endonuclease activity cleaves a single strand chromosomal DNA at the target site. Then, ORF2 with its reverse-transcriptase activity uses the free 3´OH of the nicked DNA as a primer and the mRNA as a template for LINE-1 cDNA synthesis (Cost *et al.* 2002). Retrotransposition mechanism ends with the second strand target DNA cleavage, synthesis of the second LINE-1 strand and repair of the junctions.

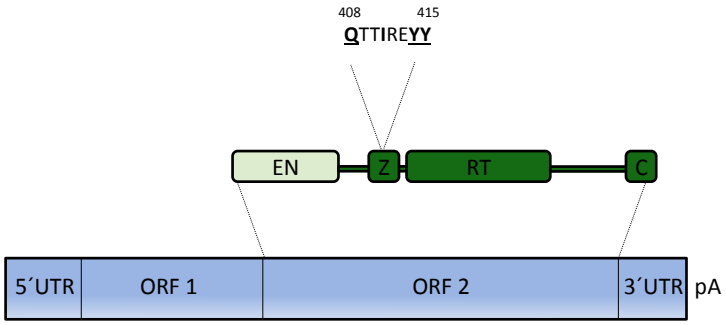


Figure 4.28. Human LINE-1 structure and PCNA interaction protein domain (PIP box)
An active human LINE-1 element typically has a 6 kb length with a 5´and a 3´untranslated regions (UTR) flanking two open reading frames (ORF 1 and 2). Downstream of the 3´UTR there is a poli A rich tail (pA). ORF 1 codifies for a 40 kDa RNA binding protein and the ORF 2 encodes a 150 kDa protein that has two main activities. At the N-terminus of the ORF 2 protein there is a domain with endonuclease activity (EN), downstream a domain with reverse - transcriptase activity (RT) and in the Carboxi termini a cysteine-rich region (C). The *PCNA interaction protein domain* (PIP box) is located in the Z region between "EN" and "RT" domains. In bold are residues implicated in the PCNA binding motif, and underlined those amino acids that are mutated to alanine in biochemical assays

LINE-1 has been reported to has low activity in somatic tissues, but more insertion frequency are found in germ-line (Kano *et al.* 2009), brain (Muotri *et al.* 2005) and even in some cancer types (Shukla *et al.* 2013). The potential negative effect of LINE-1 proliferation is controlled to some degree by diverse host mechanisms. Methylation of LINE-1 DNA controls its expression (Bourc'his 2004); small interfering RNAs generated from endogenous LINE-1 RNAs suppress retrotransposition (Yang and Kazazian 2006)

and the APOBEC3 cytidine deaminase family have been reported to also inhibit LINE-1 transposition (Muckenfuss *et al.* 2006).

Although highly autonomous, LINE-1 requires of host factor for transpose, like transcription factors (Athanikar *et al.* 2004), host RNA Pol II, and DNA repair factors. A recent study using proteomic affinity techniques characterized host proteins associated to LINE-1 ribonucleoprotein complexes (Taylor *et al.* 2013). Among identified factors is the Proliferating Cell Nuclear Antigen (PCNA), the Eukaryotic sliding clamp. Taylor *et al.* found that PCNA co-immunoprecipitates with ORF 2 complexes assembled *in vivo*, and determined that this interaction is established through a *PCNA Interaction Protein domain* (PIP box) located between the endonuclease and reverse-transcriptase domains.

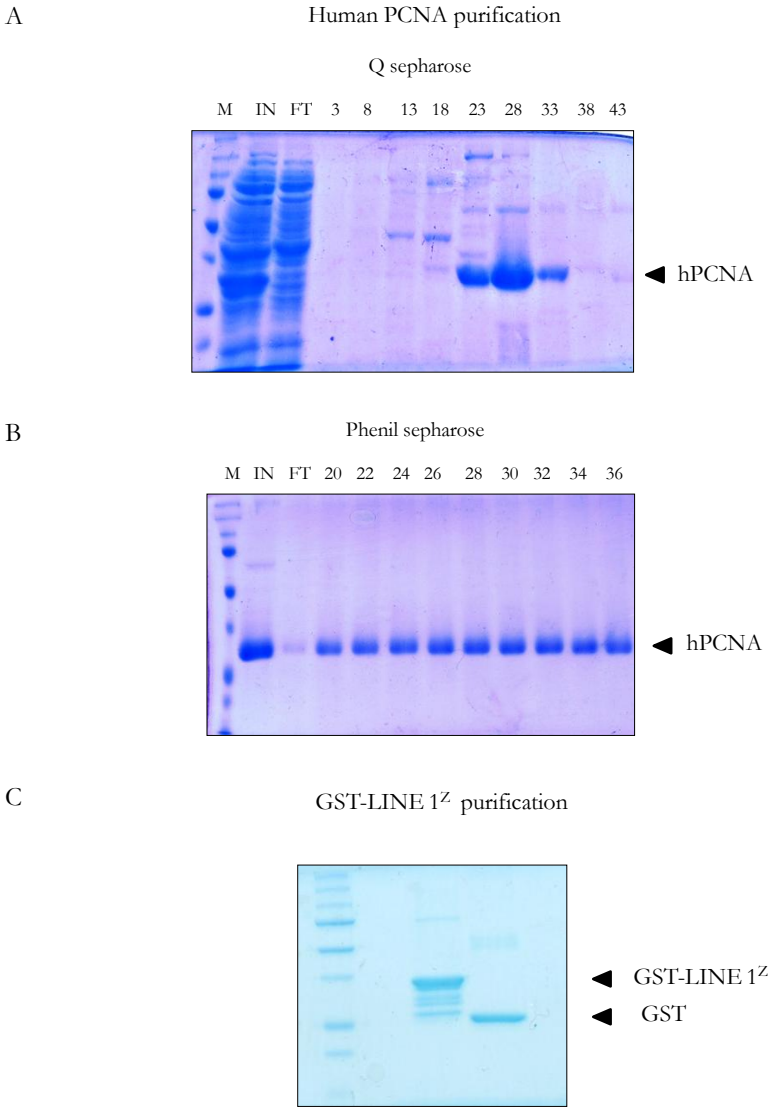


Figure 4.29. Human PCNA and GST-LINE 1Z purification.
A) PCNA was overexpressed in *E.coli* BL21 with 1mM isopropyl β -D-1-thiogalactopyranoside (IPTG). Lane 2 in the Coomassie stained SDS-PAGE gel represents the soluble protein fraction after overexpression and having been subjected to a differential precipitation with Ammonium

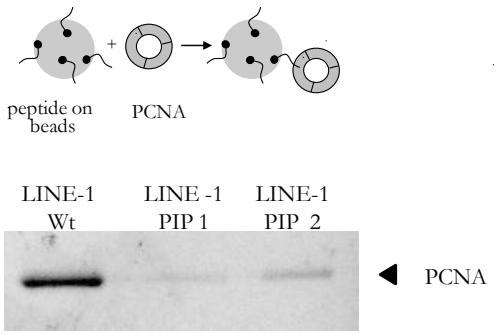
Sulfate (AS). PCNA was mainly present in this fraction that was dialyzed and applied on a Q sepharose FF (GE Healthcare) ion-exchange chromatography column (IN). Third lane is the flow-through (FT) and subsequent lanes are fractions resulted from eluting the column in a NaCl gradient (from 0.1 to 1.0 M). Fractions were PCNA was present were pooled together and dialyzed. B) PCNA was loaded on a Phenyl Sepharose FF (GE Healthcare) chromatography column (IN) and eluted with a double gradient of AS and ethylene glycol (see Methods for details). Fractions containing PCNA were pooled, dialyzed and fluorescently labeled with Alexa Fluor 350 C5-maleimide (Life Technologies). C) GST-LINE 1^{Z} and PIP box mutants were overexpressed in *E.coli* BL21 with 1mM IPTG and soluble protein fraction was applied on a Glutathione Sepharose 4 Fast Flow (GE Healthcare) resin. The column was washed and eluted with reduced glutathione. Second lane represents GST-LINE 1^{Z} wt eluted from the column and in the third lane GST is loaded for comparative purpose. M: molecular weight marker

To further characterize the LINE-1 - PCNA *in vitro* interaction and the role that PIP box residues play on it, we purified human PCNA (Fig. 4.29) and we fluorescently labeled it with Alexa Fluor 350 C5-maleimide (Life Technologies). Then we study this interaction under three different *in vitro* approaches. First, we synthesized a N-biotinylated peptide (20 aa) derived from LINE-1 sequence containing the PCNA -binding motif described in Taylor *et al.* (Fig. 4.3A). In addition we also synthesized peptides in which Q_1 of the binding motif had been changed to alanine (named PIP 1) and another one where Y_7Y_8 were also mutated to alanine (termed PIP 2) (Fig. 4.30 A). We bound theses peptides to streptavidin magnetic beads and tested their ability to retain fluorescently labeled PCNA After reactions were profusely washed, they were stopped with 1% SDS and loaded on a SDS-PAGE for results quantification. We found that Wt peptide is able to retain PCNA, meanwhile interaction was almost abolished in PIP 1 peptide and highly reduced in PIP 2 peptide (Fig. 4.30 B and C).

A

```
Wt pept     DITTDPTEIQTTIREYYKHL
PIP1 pept   DITTDPTEIATTIREYYKHL
PIP2 pept   DITTDPTEIQTTIREAAKHL
```

B



C

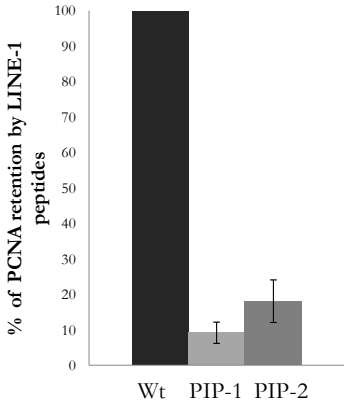


Figure 4.30. Pull down assay of PCNA by LINE-1 derived peptides containing a putative PIP box.
A) Sequences of the N-biotinylated peptides used in the binding assay. In bold are the animo acids implicated in the putative binding motif. Underlined are those residues changed to alanine in mutant versions, Q408A in PIP-1 and Y414A, Y415A in PIP-2. B) Biotinylated LINE-1 peptides were bound to streptavidin-coated magnetic beads and assayed for retention of purified fluorescently-labeled human PCNA. Reaction products are resolved in a SDS-PAGE gel and visualized on a UV-transiluminator. C) Quantification of PCNA retention as a relative percentage of the retention by LINE-1Wt peptide

Purification of ORF 2 protein entails some technical difficulties, like low purity or low purification yield (Clements and Singer 1998), therefore to further characterize PCNA - ORF 2 interaction, we designed and purified a GST fusion protein of the Z region that contains the PCNA PIP box (Fig 4.29). Additionally, we also engineered and purified recombinant proteins with the same two PIP box mutants described above. In one the glutamine in first position of the motif was mutated to alanine (named GST-PIP-1 mutant), and another one where the two tyrosines were also mutated to alanines (termed GST-PIP-2). Same amounts of purified recombinant proteins were coupled to Glutathione magnetic beads and probed to interact with fluorescently labeled PCNA. We also used GST as a

negative control, to ensure that no unspecific interaction was taking place. Results show that GST-LINE-1^{Z} retains PCNA, while GST-PIP-1 mutant does not (Fig 4.31 A). Although GST-PIP-2 mutant still have the ability to bind some of the PCNA added to the reaction, it is significantly lower than PCNA retained by GST-LINE-1^{Z} (Fig 4.31 B).

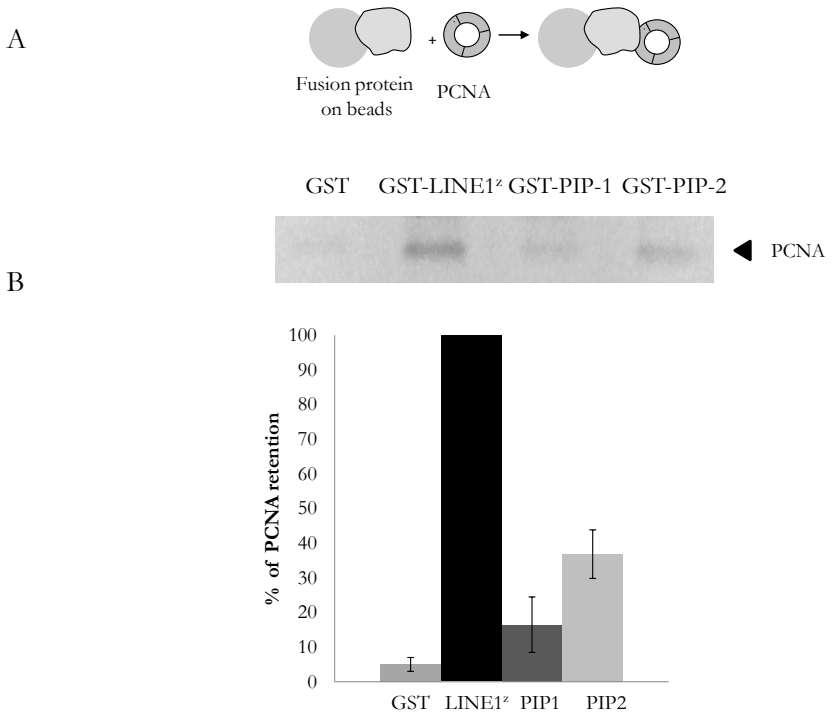


Figure 4.31 Pull down assay of PCNA by GST-LINE-1^{Z}
A) GST-LINE-1^{Z} and PIP box mutants, GST-PIP-1 and GST-PIP-2, were coupled to Glutathione-coated magnetic beads and tested for binding to labeled PCNA. Reaction products are loaded in a SDS-PAGE gel and revealed on a UV transiluminator. In first lane GST is used as a negative control. B) Quantification of results in A were done by densitometry of the bands and plotted as relative percentage of the retention by GST-LINE-1^{Z}.

Finally, a third approach to biochemically verify the direct PCNA-ORF2 interaction is using the purified recombinant protein and PCNA in a gel filtration assay. Proteins were applied on a Superdex 200 (GE) size exclusion chromatography column, fractions of 80 µl collected and loaded on a SDS-PAGE gel for visualization. GST- LINE-1^{Z} (M.W. 38,062 kDa) was loaded on the column and it eluted with a maximun peak at fraction 34. In the other hand PCNA (M.W. 28,705 KDa as a monomer) mainly eluted faster than GST-LINE-1^{Z} which is consistent with the trimer conformation of PCNA in solution (Fig 4.32 A and B). When GST- LINE-1^{Z} and PCNA are mixed together and run on the column they both co-eluted indicating that they form a stable complex (Fig 4.32 C). Furthermore,

when we calculated the pmols of both proteins in each fraction, stoichiometry suggests that one molecule of GST- LINE-1^{Z} is interacting with a trimer of PCNA.

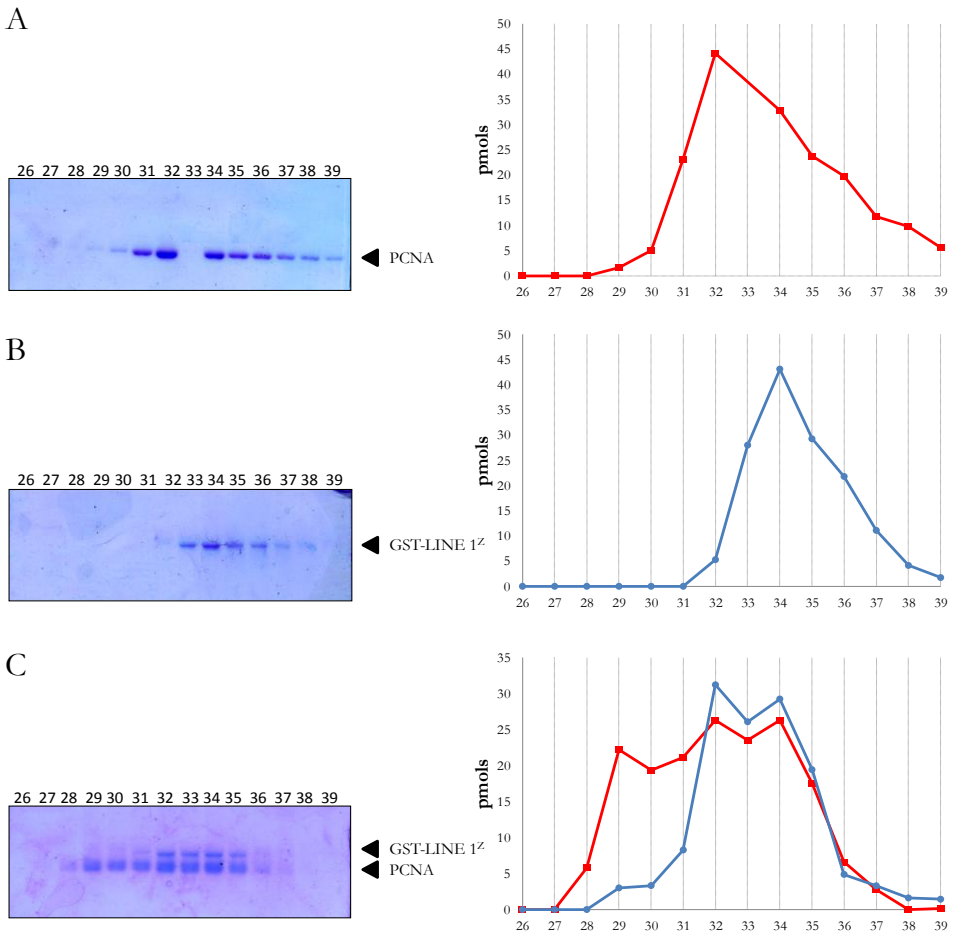


Figure 4.32. FPLC assay. PCNA interacts with GST-LINE-1^{Z}
PCNA (86 kDa as a trimer) and GST-LINE-1^{Z} (28 kDa) are independently loaded on a size exclusion chromatography column and eluted fractions analyzed on Coomassie stained SDS-PAGE gels (left panels in A and B respectively). Fraction number are indicated above each gel (note that fraction 33 is missed for PCNA). Right panels represent pmols of protein in each fraction C) PCNA and GST-LINE-1^{Z} are mixed, loaded on the column and eluted fractions visualized in a SDS-PAGE gel (left panel) revealing that they co-eluted as a complex. pmols of both proteins in each fraction are plotted (a red line for PCNA and a blue one for GST-LINE 1^{Z} in the right panel) and results suggest that one mol of GST-LINE 1^{Z} interacts with one trimer of PCNA.

Taken together these results indicate that LINE-1 interacts with PCNA through the PIP box located in the Z region of ORF2 protein as previously described (Taylor *et al.* 2013). Moreover, our date suggest that the glutamine residue in first position of the binding motif play a critical role in the interaction, because an alanine substitution in this position knock out the interaction (Fig 4.30 B and Fig 4.31 A).

Discussion

5.1 The source of IS orientation biases in chromosomes

The main hypothesis guiding our analysis of ISs in bacterial genomes was that, if IS transposition events are associated with host replication, they could present orientation patterns in the chromosome. Importantly, our study was severely limited by various factors, namely 1) the heterogeneity present within some IS families (e.g., variability in the orientation of the transposase gene with respect to other elements within the IS); 2) the requirement of a relatively high number of ISs per chromosome to achieve statistical significance (orientation patterns in IS families with low copy number per chromosome could be undetectable); 3) our inability to distinguish between IS insertions resulting from transposition within the chromosome from those incorporated into chromosomes within large blocks of DNA ("genomic islands," prophages); and 4) the uncertainty derived from using current GC skew as a proxy of replication fork orientation, as any chromosomal rearrangements would tend to randomize any orientation bias.

Despite the mentioned limitations, our analysis of IS orientation in bacterial chromosomes revealed strong orientation bias ($P<10^{-2}$) for three IS families in Proteobacteria, two in Actinobacteria, and ten in Firmicutes (Table 4.1). What could be the underlying biological phenomenon generating a biased orientation of ISs in chromosomes? Biases could have been generated by 1) preferred insertion of ISs in non-randomly oriented sequences in the chromosome, 2) post-insertion selection favoring specific orientation, or 3) by transpososome interaction with an asymmetrical structure within the replication fork. The first possibility, target sequence specificity, has been observed for Tn*7*, in which the Tn*7*-encoded protein TnsD directs insertions to a specific location on the chromosome near Ori (attTn*7*) (Waddell and Craig 1988). Also IS*110*, a family for which we find strong orientation bias in Proteobacteria and Firmicutes, could possibly reflect oriented insertion into targets such as REP sequences (Tobes and Pareja 2006), the terminal repeats of IS*21* (Partridge and Hal,l 2003), or the recombination sites (attC) of integron gene cassettes (Tetu and Holmes 2008; Post and Hall 2009), which could themselves be biased. However, most ISs show little or weak sequence specificity (Chandler and Mahillon 2002), and the highly distributed placement of most ISs in chromosomes of phylogenetically diverse bacteria renders unlikely the possibility of generalized sequence targeting as a source of bias for most IS families. The second mechanism, post-insertion selection, could possibly generate a bias if, for example, transcription from upstream genes altered expression of the transposase and this had an effect in viability or if the IS altered regulation of neighboring genes (Plague 2010). However, our analysis of orientation of ISs in bacterial chromosomes

does not reveal any global orientation bias of the IS population in chromosomes, even in those with a very strong gene orientation bias, such those of Firmicutes (Fig. 4.3). Finally, an interaction between transposition and replication could also generate an orientation bias, as it has been recently described in detail for two IS families: Tn*7* (Parks *et al.,* 2009) and IS*200* (Ton-Hoang *et al.,* 2010). In the case of Tn*7*, an accessory factor, TnsE, interacts physically with the β sliding clamp and targets the Tn*7* transposase preferentially to conjugative plasmids. A study of the orientation of many independent chromosomal insertion events revealed a clear TnsE dependent, replication-dependent bias (Peters and Craig 2001). On the other hand, IS*200* shows a clear orientation bias in bacterial chromosomes, explained by its requirement for ssDNA found mainly in the lagging strand at replication forks (Ton-Hoang *et al.,* 2010). However, it is important to note that interaction with replication does not necessarily impose an orientation bias (see later). Our observations (Fig. 4.1, Table 4.1) suggest that there is a replication-dependent bias in some IS families and that whether the bias is in favor or against the direction of movement of the replication fork is IS-family dependent. For many families, this bias is strong in Firmicutes but absent in Proteobacteria and Actinobacteria. Because it is unlikely that ISs are mechanistically different in Firmicutes, when compared with the other phyla, this result strongly suggests that ISs insertions in Firmicutes behave differently than in other phyla as direct consequence of distinct chromosomal replication dynamics in this group.

5.2 Sliding clamp as a general link between transposition and replication

In our search for host replication factors that interact with transposition, we performed a systematic search for the β binding motif in transposases. Then, we assayed synthetic peptides derived from *E. coli* transposases, containing the putative binding motif, for the interaction with *E. coli* β clamp. Our approach was limited by 1) our ability to recognize the canonical β interaction motif in *E. coli* transposases, as variation within the motif is high, even in well-characterized enzymes (Dalrymple *et al.,* 2001), 2) the sensitivity of the biochemical techniques, and 3) the absence of some IS families in sequenced *E. coli* genomes. However, we found the motif for interaction with β in nine different transposase families. Furthermore, we biochemically probed that those transposases interact with *E. coli* β through the identified motif. Although the β -binding motif is short and relatively poorly conserved, the competition assay with the strongly binding ligand PolIV^{LF} assured that the proposed peptides bind to β. Besides, this experiment mapped the interaction to the canonical hydrophobic pocket in β where all other enzymes, included PolIV, also bind (Fig.

4.7). All together suggest a possible general mechanistic link between transposition and chromosomal replication.

β provides processivity to DNA polymerases, but the main role of sliding clamps in most studied systems, such as Okazaki fragment processing or DNA polymerase switching during lesion bypass, is targeting enzymes to active replication sites to couple and coordinate their activities. Most of transposition mechanisms require DNA repair or replication by a DNA polymerase, so β would allow polymerase recruitment, because the five DNA polymerases present in *E. coli* require β (López de Saro *et al.,* 2003). β could be bound by the transposase first to initiate the transposition reaction and then used to target the appropriate polymerase in a subsequent step. However, and given the diversity of transposition mechanisms, it is possible that β is used in distinct ways by the different transposases. It has been proposed that β targets Tn*7* to replication in conjugative plasmids as a mechanism for dissemination to new hosts (Peters and Craig 2001; Parks *et al.,* 2009). In the case of IS*200*, binding to β could help the transposase to localize to sites with increased amounts of ssDNA, such as replication forks or repair sites, thus increasing the efficiency of the excision or insertion processes.

Our results indicate that IS families with diverse transposition mechanisms (DDE, HUH and S-transposases) could interact with the replisome similarly, suggesting convergent evolution for interaction with the host. For example, β binding motifs are found in transposases with different mechanisms like Tn*3* (co-integration), IS*91* (rolling circle) or IS*30* (copy-paste). Similarly, a putative motif can be found in the C-terminus of OrfB of IS*200*/IS*605* in Cyanobacteria (Transposase alignments in Fig. I.4) but seems absent in other phyla at that position and variants of Tn*7* harbor β binding motifs in proteins TnsC or TnsE (Parks *et al.,* 2009).

Transposase sequences show a considerable degree of diversity and degeneracy even within the same IS family. It is tempting to speculate that given the relatively simplicity of the β binding motif, which often but not exclusively, resides in unstructured sequences at the C-termini , could be easily not only deleted but also be regenerated *de novo* from unrelated sequences by random mutation and selection. This could explain the appearance of β motifs at distinct locations in homologous transposases. For example, strong β binding motifs can be found in TnpB of IS*66*a, in two different positions within TnpC, or in both proteins (Fig. 5.1). Likewise the secondary motif identified at the C-terminal sequence of *Acidithiobacillus ferrivorans* IS*1634* also supports this idea (Fig. 4.23). This pattern could also explain why binding motifs are found within apparently different

sequence context in members of the same IS family. This is the case of transposases of IS*200*/IS*605* (OrfB) in Cyanobacteria or β-motif sequence alternatives described in IS*200* of *E. coli* (Transposase alignments in Fig. I.4).

In eukaryotic organisms, PCNA-binding has been detected for Pogo, a Drosophila transposase (Warbrick *et al.,* 1998; Warbrick 2000) and, recently, for the endonuclease/reverse transcriptase of LINE-1, where it has been shown to be a requirement for retrotransposition (Taylor *et al.,* 2013). We have studied *in vitro* the role that play critical residues of the PCNA interaction protein domain located in LINE-1, and show how a single amino acid substitution could dislocate the interaction.

Collectively, these findings suggest that the ability to interact with sliding clamps of prokaryotic and eukaryotic organisms is likely to have evolved repeatedly and independently in the different transposase families and even within the different lineages in the same family. Hence, this represents an extraordinary example of evolutionary convergence of far related IS families to interact with sliding clamps. Evolutionary convergence has also been proposed for transposase domains and transpososome architecture (Montaño *et al.,* 2012).

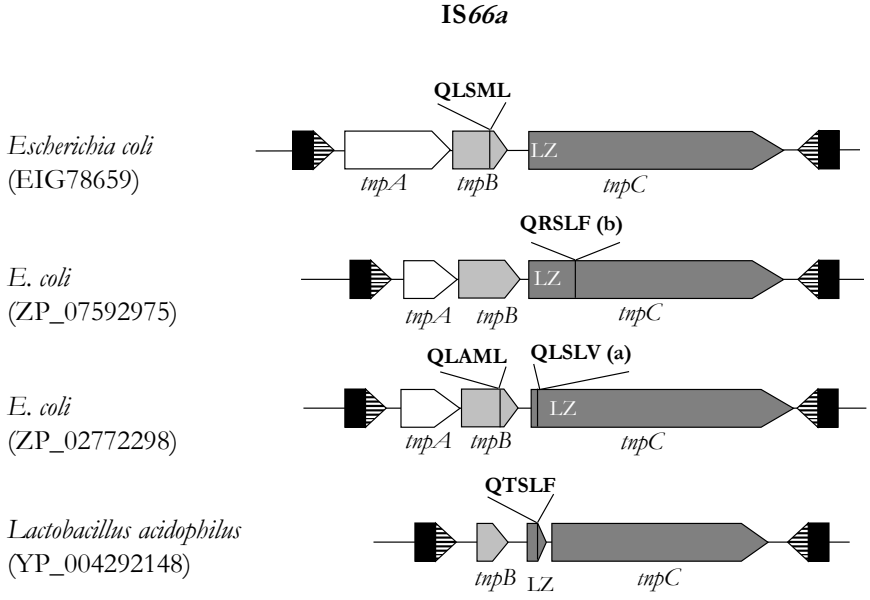


Figure 5.1. Structure of IS*66* and diversity in sequence and location of β -binding motifs. The motif can be present in TnpB, TnpC, or in both. In TnpC, it can be located upstream or downstream of the leuzine zipper domain (LZ). In some Bacilli (Firmicutes), the LZ domain is an independent open reading frame and presents the β -binding motif at its C-terminus

5.3 Replisome composition could explain the IS orientation biases in Firmicutes

Although β interacts with transposases involved different transposition pathways, this alone does not imply the generation of an orientation bias for these IS families in chromosomes. However, the strong orientation bias of some IS families found in Firmicutes (Table 4.1) could be readily explained by three concurring circumstances: first, the interaction of asymmetric transpososomes with β (symmetric transpososomes would not result in chromosomally biased orientations); second, the fact that β is loaded on DNA in a regular, oriented, manner by the replisome and that all factors that interact with it do so on the same face of the ring (López de Saro 2009); and third, differences in the amount of β associated with the synthesis of the leading and lagging strands (Fig. 5.2 C). In *B. subtilis*, β is slowly recycled after synthesis of the Okazaki fragments (lagging strand) and tends to accumulates in highly condensed "clamp zones," where β is presumably free to interact with other factors (Su'etsugu and Errington 2011). In the other hand, in the *E. coli* replisome, β also accumulates during lagging strand synthesis, but stoichiometry studies have proposed that β molecules that remain bounded to the DNA are by far, less abundant than in Firmicutes clamp zones (Moolman *et al.,* 2014) The strongly asymmetric content of β associated with synthesis of the leading versus lagging strands in *B. subtilis* could explain the orientation bias found for IS families in Firmicutes (Fig. 5.2 C).

In Proteobacteria, however, we find three strongly biased families, IS*91*, IS*110*, and IS*200*. Although we have found an interaction of IS*91* and IS*200* with β, other additional mechanisms could add to their orientation pattern. IS*91* uses a rolling-circle mechanism that requires DNA synthesis and that, because the IS ends are different, is strongly asymmetric (Garcillán-Barcia *et al.,* 2001; Curcio and Derbyshire 2003; Chandler *et al.,* 2013). IS*200* orientation is determined by its use of ssDNA and preferential insertion in the lagging strand (Ton-Hoang *et al.,* 2010). No mechanistic model is available for IS*110* that could explain its orientation, but, as mentioned before, specific targeting could also be involved.

Our model relies critically in three levels of asymmetry generating the observed biases: binding to β (Fig. 5.2 A), the transpososome (Fig. 5.2 B) and the replication fork (Fig. 5.2 C). Although asymmetries derived from the first two have been analyzed extensively, only a few transpososomes have been studied in structural detail (reviewed in Dyda *et al.,* 2012). Although transpososomes consist of homomultimeric transposases, major conformational and functional asymmetries (e.g., sequential cleaving of DNA ends) have been found, for example, in the transposition pathways of Tn*5* (Reznikoff 2008), Mu

(Montaño *et al.,* 2012), IS*91* (Garcillá n-Barcia *et al.,* 2001), IS*3* (Sekine *et al.,* 1999), or IS*200* (Ronning *et al.,* 2005). In all cases, if β binds preferentially to one of the transposases, then an orientation bias during insertion on DNA could be the result (Fig. 5.2 B). Otherwise, if interaction of β with either transposase is identical, the result would be random orientation with respect to DNA. Both possibilities are plausible, and detailed interaction studies will be required to study what is the case for each transpososome architecture. In the well-studied IS*200* transpososome, an ssDNA transposition system, the architecture is inherently asymmetric due to the polarity of ssDNA. According to our peptide data (Fig. 4.6), the region of interaction of IS*200* TnpA with β would align with an α-helix close to the catalytic tyrosine, as revealed by the IS*Hp608* crystal structure (Ronning *et al.,* 2005). This C-terminal α-helix is likely to be highly flexible, but it is uncertain to what extent an interaction of either subunit with β would impose an additional asymmetry to the complex.

We have been unable to identify a β motif in 12 IS families that show orientation bias in one or more Phyla. A strong possibility is that we have failed to detect the β motif in sequences from these families, as its conservation is weak and transposases show high variability. Other possibilities are that these transposases bind β using a noncanonical motif or interaction with other replication structures or host factors.

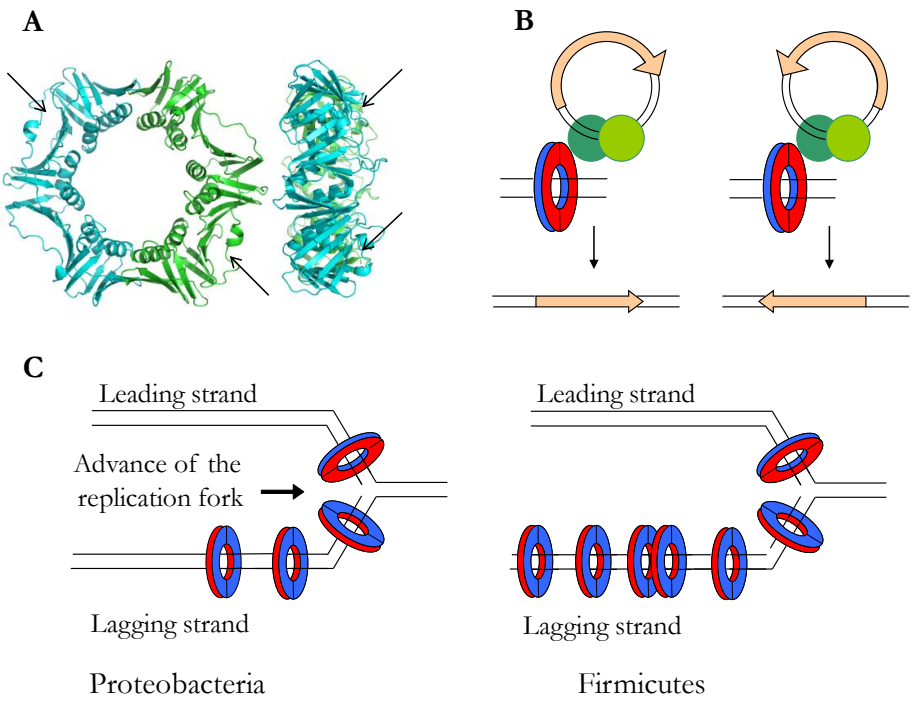


Figure 5.1 Structural and functional asymmetries contributing to biased orientation of ISs in chromosomes.
A) Structure of *Escherichia coli* β, front (left) and side (right) views (PDB: 2POL). Arrows indicate the hydrophobic pockets on the surface of each monomer of β that are the sites of interaction of all β partners studied and of all the transposase peptides described in this study. B) The asymmetry of transpososomes (green circles) in their interaction with β could determine the orientation of the transposase gene (orange arrow). The interaction "face" of β is colored red, the other blue. C) Models of replisomes of *E. coli* and *Bacillus subtilis*. β is loaded on DNA by the γ-complex, which for leading strand synthesis positions β facing the direction of movement of the replication fork and in the opposite orientation in the lagging strand. On the right panel, β accumulates in "clamp zones" as the *B. subtilis* replisome progresses, possibly due to slow recycling after Okazaki fragment synthesis, creating an asymmetry in the distribution of β associated to the synthesis of leading and lagging strands (Su'etsugu and Errington 2011). On the left panel, *E. coli* replisome shows a less condensed "clamp zone" β associated with the synthesis of lagging strands (Moolman *et al.,* 2014).

5.4 Transposase interaction with β. Implications in transposition self-regulation

Transposable elements may have detrimental effects in the host genome integrity if they expand uncontrollably. In fact, transposition is subjected to tight regulation imposed both by the host and by intrinsic IS mechanisms (Nagy and Chandler 2004), which are reflected in the low transposition rates commonly observed in nature. The widespread distribution of the interaction between phylogenetically distant transposases and sliding clamp, suggests that β binding could be a fundamental aspect of transposition regulation. To study whether this interaction is maintained between β and the transposase (Tnp) of the well-known Tn*5* transposon, we used purified proteins and demonstrated that Tnp is also able to bind β. We mapped the interaction and found that a 7 amino acids N-terminal deletion fails to bind β. However, we did not identify a canonical β binding motif in those 7 amino acids (Dalrymple *et al.,* 2001). This could be explained under three different considerations. First, due to the high diversity and loose conservation of binding motifs found in transposases, we have not succeeded in the identification of key amino acids that define the binding motif. Another plausible possibility is that the N-terminus of Tnp contains a noncanonical binding motif. Finally, we cannot rule out the possibility that the binding motif is in another region of the protein, and that a 7 amino acids N-terminal deletion could induce a conformational change which masks the putative β binding motif elsewhere in the protein.

The involvement of the N-termini of Tnp in the interaction with β, an essential replication factor, correlates with the previously described citotoxic effect of Tnp (Weinreich *et al.,* 1994a). It has been demonstrated that the toxicity of Tnp is not dependant of the transposition activity nor DNA binding, and relays on few N-terminal amino acids. In fact, deletions of 3 to 11 amino acids in the N-termini of the protein suppress the lethal effect. Moreover, it has been proposed that the recruitment of some

replication host factor by Tnp for its own transposition activity, could be the source of the toxicity (Weinreich *et al.,* 1994a). β is limiting in the cell and is an essential factor for host enzymes like polymerases or repair factors. Thus, our findings that Tnp binds β, and that amino terminal deletions in Tnp that relieve *Tn5* toxicity also prevent β binding, strongly suggest that in the cell, Tnp is likely competing with other host enzymes for β, explaining the described toxic phenotype of Tnp. The fact that all transposases found interacting with β do so at the same position in competition with replication and DNA repair factors predicts that an excess of ISs in the genome or transposases with strong β binding motifs could be disruptive to DNA replication.

We have also observed that when incubating DNA with Tnp in the presence of β, a new DNA complex is formed. Tnp mainly binds DNA through the N-terminal region of the protein (Weinreich et al 1994b). However, it has been suggested that C and N-terminus of Tnp establish close contacts in solution, limiting Tnp ability to bind DNA and to form synaptic complexes (Reznikoff 2008). Thus, Tnp should undergo a sort of conformational changes that allow the enzyme to effectively bind DNA and transit from a monomer to a dimer state. The fact that the N-terminus of Tnp is also involved in the interaction with β, suggests that Tnp is tethered to DNA through contact with the β clamp. Tnp structural self-regulation may be loosen by β binding which could promote the necessary conformational change to expose the Tnp DNA binding region and the dimerization domain. Interaction with the sliding clamp also targets Tnp to the replication fork, so it could co-localize with host repair and replication factors required in the transposition mechanism. Besides, Tnp coupled to the sliding clamp and the movement of the replication fork, may also enlighten how Tnp search for outside ends sequences of the transposon along the chromosome.

The self-inhibition state of Tnp and its low activity have led to the development of a wide collection of random Tnp hyperactive mutants. However, described mutant induces a modest increase of around 10-fold in transposition activity *in vivo* (Weinreich *et al.,* 1994c; Wiegand and Reznikoff 1992). Greater activities are achieved when selected mutations are used in combination (Weinreich *et al.,* 1994c). We have reported that mutations in the conserved L363 or L366 (Reznikoff *et al.,* 2004) highly increase observed transposition events *in vivo* over 100-fold compared to Tnp. Moreover, both mutations promote to some extent, increased ability to bind DNA. Since no X-ray crystal structure of the monomeric Wt Tp are available, is difficult to predict what kind of structural change could induce our mutants. On the other hand, when the crystal structure of the synaptic complex is

examined (Davies *et al.,* 1999), mutated leucines are in the vicinity of both the tryptophan residue 450 and N terminal amino acids (Fig. 5.3). W450 residue is believed to be implicated in hydrophobic and likely aromatic contacts between N and C termini contributing to Tnp self-inhibition. In fact mutants W450C and W450R slightly increase transposition *in vivo* and DNA binding while W450F has similar phenotype to control Tnp (Gradman and Reznikoff 2008). One could speculate that L366F may replace the aromatic contacts between W450 and the N-terminus, and thus release the inhibitory interaction between them. Another explanation could rely in conformational changes introduced by L363A or L366F that also dislocate the N and C inhibitory contacts. Nonetheless, clearly more experimentation is required to understand the biochemical mechanisms behind our described hyperactive mutants and to determine whether β induced DNA complex is a fully functional transpososome.

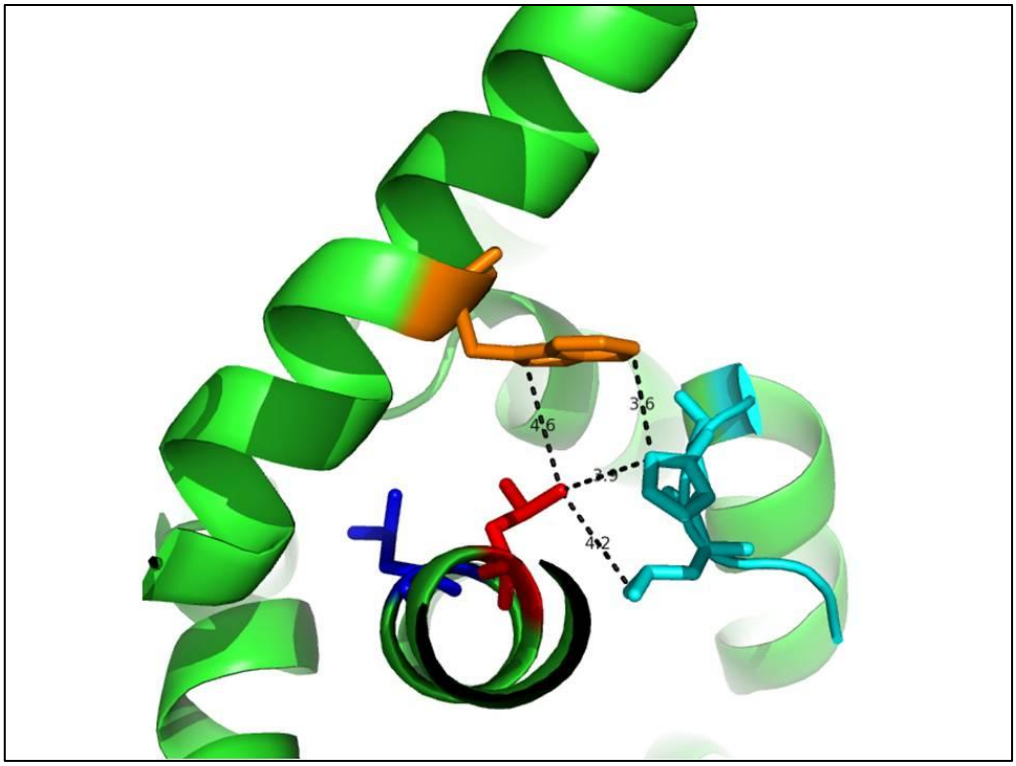


Figure 5.3. Detail of the crystal structure of Tn*5* synaptic complex.
8 N terminal amino acids are colored in cyan. Lateral chain of L6 and H7 (cyan), L363 (blue), L366 (red) and W450 (orange) are shown. Illustrative atomic distance (in Å) between closest atoms of L366 - L6 (4.2), L366 -H7 (3.9), L366 - W450 (4.6) and W450 - H7 (3.6) are represented by dotted lines

5.5 IS proliferation and HGT. Role of transposase interaction with sliding clamps

Another main objective of this work was to identify an active insertion sequence from a natural environment and explore its interaction with the host β sliding clamp regarding its ability to proliferate within the chromosome and its potential for dispersal to other species.

Our 600-generation laboratory evolution experiment with *Acidiphilium* sp. PM, aimed at detecting changes in IS copy number that became fixed in the culture by comparative microarray hybridization, found three cases of proliferation and six deletions. Other IS transposition events could have occurred and remained undetected because they were detrimental to the host and were selected against, because they were lost by genetic drift, or because they involved IS relocation in the chromosome (in the case of cut and paste transposition). As it is the case with other studies that have analyzed spontaneous global transposition activity in genomes, our study suggests that only a small fraction of the ISs detected by sequencing are active (Martusewitsch *et al.,* 2000). The detectable transposition activity can vary greatly with IS element and strain (from 10^{-3} to 10^{-7} per generation, according to some estimates) (Kleckner 1990; Reznikoff 2008; Sousa *et al.,* 2013). Unlike the well-defined and relatively predictable point mutation rates resulting from the combined action of DNA polymerases and repair genes, transposition rates are likely the result of the combination of diverse factors such as IS sequence variation, transposase expression and activity, chromosomal location effects, or various host regulatory mechanisms. Additionally, it is likely that the process of isolation and adaptation to laboratory conditions change transposition rates. For example, the observed transposition events in *Ferroplasma* were substantially more frequent in culture than in environmental samples of the same organism (Allen *et al.,* 2007).

We also explored the expression of transposases at the initial point of our long-term culture, and found a clear correlation between transcribed transposases and those in which copy number changes were detected. Although transposase gene expression under different laboratory conditions or induced stress situations have not been study in detail, metatranscriptomic and metaproteomic analysis of bacterial populations have also detected relatively quick changes in transposase gene expression in response to environmental stimuli (Hewson *et al.,* 2009; Mueller *et al.,* 2010). However underlying mechanisms of global IS expression are unknown.

Because ISs are subject to inactivation by mutation, horizontal transfer is an essential aspect of their lifestyle and critical for the persistence of ISs in bacterial populations (Cerveau *et al.,* 2011b; Bichsel *et al.,* 2010). We studied whether transposase

interaction with β could be a strong barrier for IS exchange among phylogenetically-distant but habitat-sharing organisms (Wagner and de la Chaux 2008; Hooper *et al.,* 2009). We found that *Acidiphilium* IS*1634* Tnp, that we have reported to be active in our long-term culture, is not only able to bind to *Acidiphilium* β but also binds to *Leptospirillum* or *Escherichia* β , and even to an archaeal sliding clamp PCNA from *Methanosarcina*. We also mapped the interaction to a binding motif in the C-termini of Tnp (Fig. 4.21). Although β and PCNA share no detectable sequence similarity, these proteins are structurally and functionally very similar, and the binding motifs of the proteins interacting with them are highly related (Dalrymple *et al.,* 2001; López de Saro 2009; Hedglin *et al.,* 2013). In fact a peptide derived from *Methanosarcina* IS*1634* Tnp can bind both PCNA and *Acidiphilium* β (Fig. 4.25). Similarly, a *Methanosarcina* IS*200* derived peptide interacts with PCNA and *E. coli* β (Fig. 4.8). Our results suggest that transposases could interact relatively easily and bidirectionally with the replication machinery of bacterial or archaeal hosts. On the other hand, because sliding clamps are universal and highly conserved, adaptation of transposases to binding β in new organisms could require only subtle sequence changes, facilitating IS exchange among phylogenetically distant organisms.

With a few exceptions (e.g., ISH6, found only in Archaea), most IS families can be found in bacterial and archaeal genomes, and genome sequencing suggest that movement of these mobile elements between the two domains is fluid (Filée *et al.,* 2007). However, a recent survey of prokaryotic IS elements in eukaryotic genomes detected few events of IS transfer in recent eukaryotic evolutionary history, with the possible exception of cyanobacterial IS*607*(Gilbert and Cordaux 2013). Although HGT events from prokaryotic to eukaryotes have been documented extensively (Hotopp *et al.,* 2007; Schönknecht *et al.,* 2013), it remains to be investigated if mobile elements are involved.

It is remarkable that all enzymes involved in mutagenic processes, namely the DNA polymerases, the mismatch repair system, and transposition, directly interact with the β sliding clamp. Indeed, a recent experimental study demonstrated a direct 'conflict' between these processes, as mismatch repair mutator alleles, often present in bacterial populations, limit insertion sequence proliferation in the early stages of invasion in a new host (Fehér *et al.,* 2012). The relation between IS and host genomes has been described as a dynamic equilibrium state, which may become transiently perturbed by the rapid expansion of IS populations and that can be explained just with two parameters: the duplication-deletion ratio and the HGT-deletion ratio (Iranzo *et al.,* 2014). We propose that the ability of transposases to interact with β is one of the molecular processes that contribute to the

"duplication" component of the first parameter and, as such, it is subject to natural selection. Variations in the binding site that increase the affinity for β may be selected upon colonization of new hosts and cause explosive proliferation of IS. On the contrary, variations that decrease the affinity for β would reduce transposition rates, contributing to a reduction in IS populations. Future studies will be required to understand how these various processes contribute modulate genetic variability and to the adaptive potential of bacterial populations.

5.6 Transposase affinity to β sliding clamp influences transposition rate

The β binding motif is present in a large number of proteins involved in DNA synthesis and repair. Since β is limiting in the cell (López de Saro 2009), the concentration of transposases and the strength of their interaction could determine the efficiency of transposition. All proteins so far studied in detail interact with β on the same face of the ring, binding competitively to the same conserved 'hydrophobic pocket'. Using peptides, we have shown that different families of transposases also bind competitively to the same site on β, including transposase of several members of the IS*1634* family. Moreover, we have also determined that different binding motif sequences, even within the same IS family, have diverse β affinity (Fig. 4.9, Fig 4.10 and Fig. 4.21). The affinity with which enzymes bind β is likely to be finely-tuned in the cell to accommodate the various processes for which β is an essential component (Maul *et al.,* 2007; López de Saro 2009)

The genome ecosystem hypothesis (Kidwell and Lisch 1997; Brookfield 2005) suggests that mobile elements in a genome are analogous to an ecological community in which its components have a limited access to host resources (e.g., space in the chromosome, host factors required for transposition). Their fate would be a function of their ability of adaptation and proliferation in a given genomic environment including the interaction with β. Our *in vivo* transposition assay with IS*1634* aimed to determine if the relative affinity of a transposase for β could alter its chances of success. Results show that transferring IS*1634* from *Acidiphilium sp. PM* to *Escherichia coli* is greatly favoured by point mutations in the motif that make the interaction stronger (Fig. 427). The designed IS*1634* transposase CN mutant defines a new category of hyperactive transposases: one with improved interface with the host. The search for hyperactive transposases has been based typically in genetic screens of random point mutants which are then combined in a single transposase to greatly increase its activity. For example, commercial Tn5 transposase contains three amino acid changes which increase activity by > 100-fold (Reznikoff 2008),

PiggyBac mutations in seven amino acids generate a 17-fold increase in activity (Yusa *et al.,* 2011), and a combination of multiple mutations in Sleeping Beauty can generate up to a 100-fold increase in transposition (Mátés *et al.,* 2009). The principles derived from our designed hyperactive IS*1634* mutant could potentially open the door for development of novel hyperactive transposases for use in biotechnology, or for the design of transposases with an expanded range of hosts by designing a binding motif *ad hoc* for the host.

5.7 Final remarks

By interacting with β, transposases are targeted to the replication fork, thus associating chromosomal replication with transposition. Coupling of transposition to replication would ensure, first, the possibility of recombination and repair with the sister chromosome, which would limit the potential damage to the host caused by IS excision and the generation of double-strand breaks. Second, IS elements with a conservative (cut and paste) mechanism of transposition would have a chance of jumping to the sister chromosome and therefore increase their number in that replicon. Third, transposition would co-localize with repair factors required in the later stages of transposition to fill in the gaps and ligate (e.g., both DNA polymerase I and DNA ligase interact with β (López de Saro and O'Donnell 2001)) Fourth, transposases could couple β-binding with allosteric changes that initiate transposition, such as monomer-dimer transitions, binding to the end sequences of the IS or be released from a state of self-inhibition which has often been observed in these enzymes (Reznikoff 2008). Formation of the transpososome involves dimerization and conformational changes which often involve C-terminal regions of the protein (Braam *et al.,* 1999; Dyda *et al.,* 2012) which are, usually, also the sites of interaction with β. Finally, since sliding clamps are universal and highly conserved among species, transposase interaction with β or PCNA will ensure a rapid integration within the replication machinery of a new host allowing IS proliferation. Furthermore, increasing the affinity for β could cause explosive proliferation of IS in the chromosome and conversely, mutations that blur the binding motif, may diminish the transposition ability of ISs. Using sliding clamps as a dispersal platform could provide the key to the ubiquitous nature of IS in prokaryotic genomes, highlighting a remarkable example of extreme molecular adaptability.

Conclusions

The results of this Thesis have lead to the following main conclusions:

1. Up to 18 families of insertion sequences (IS) in some bacterial groups show significant orientation bias in their insertion patterns in chromosomes. These orientation biases are related with the movement of the replication fork and not derived from post-insertion selection.

2. We found a widespread motif among transposases for interaction with the β sliding clamp, an essential host replication factor. We demonstrated that transposases belonging to up to 10 IS families can bind to the *E. coli* β clamp *in vitro*. Binding occurs via a short conserved motif usually located at the C-terminal domains of the transposases and competitively with other β-interacting proteins.

3. The interaction of transposases with the β clamp, the asymmetry of transpososomes, and the phylum-dependent β distribution in replisomes, could contribute to the observed orientation bias of ISs in chromosomes. Since transposase families are non-homologous, their common interaction with the sliding clamp demonstrates a case of evolutionary convergence.

4. The Tn*5* transposase interacts with the *E. coli* β clamp. The interaction likely releases the transposase from a state of self-inhibition, thus promoting DNA binding. We also discovered and characterized two novel hyperactive mutants of Tn*5* transposase.

5. We identified an active transposase in a natural environment that could bind to β from various species, some of them distant phylogenetically. The interaction between transposases and sliding clamps can be conserved and function over large phylogenetic boundaries. It can even be extended to archaeal transposases and sliding clamps, explaining the fluid exchange of insertion sequences between Bacteria and Archaea.

6. Optimization of the interacting site on the transposase to fit the consensus β-binding motif results in a transposase with increased transposition rate, opening a possibility for the rational design of hyperactive transposases for use in biotechnology, or for the design of transposases with an expanded range of hosts.

Los resultados de esta tesis han dado lugar a las siguientes conclusiones principales:

1. Hasta 18 familias de secuencias de inserción (SI) en algunos grupos bacterianos muestran un sesgo significativo en sus patrones de inserción en los cromosomas. Estos sesgos de orientación están relacionados con el movimiento de la horquilla de replicación y no son derivados de selección post-inserción.

2. Encontramos entre las transposasas un motivo ampliamente distribuido para la interacción con β *sliding clamp*, un factor esencial de la replicación del hospedador. Demostramos que las transposasas pertenecientes a 10 familias de SI pueden interaccionar in vitro con β *sliding clamp* de *E. coli*. La interacción ocurre a través de un motivo corto y conservado localizado generalmente en el dominio C-terminal de las transposasas que compiten frente a otras proteínas que interaccionan con β.

3. La interacción de las transposasas con β, la asimetría de los transpososomas y la distribución de β en los replisomas dependiente del phylo, pueden contribuir al sesgo en la orientación de las SI observado en los cromosomas. Al no ser homólogas las familias de transposasas, su común interacción con el *sliding clamp* demuestra un caso de convergencia evolutiva.

4. La transposasa de Tn*5* interacciona con β *clamp* de *E. coli*. La interacción posiblemente relaje el estado de auto-inhibición de la transposasa, promoviendo la unión a DNA. También hemos descubierto y caracterizado dos nuevos mutantes hiperactivos de la transposasa de Tn*5*.

5. Hemos identificado una transposasa activa en el medio natural que puede unirse a β de varias especies, algunas de ellas distantes evolutivamente. La interacción entre transposasas y *sliding clamps* se puede conservar y ser funcional a través de grandes límites filogenéticos. Esta interacción se puede incluso extender a transposasas y *sliding clamps* de arqueas, explicando el fluido intercambio de secuencias de inserción entre bacterias y arqueas.

6. La optimización del sitio de interacción en la transposasa para acomodarlo al motivo consenso de unión a β, resulta en una transposasa con una tasa de transposición incrementada, abriendo la posibilidad de un diseño racional de transposasas hiperactivas para el uso en biotecnología, o el diseño de transposasas con un rango de huéspedes ampliado.

References

Allen E.E., Tyson G.W., Whitaker R.J. Detter J.C., Richarson P.M. Banfield J.F. (2007). Genome dynamics in a natural archaeal population. Proc. Natl. Acad. Sci. USA. 104, 1883-8.

Altsch S.F., Madden T.L., Schäffer A.A. Zhang Z., Miller W., Lipman D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.

Amaral-Zettler L.A., Zettler E. R., Theroux S.M., Palacios C., Aguilera A., Amils R.(2011). Microbial community structure across the tree of life in the extreme Rio Tinto. ISME J. 5, 42-50.

Barabas O., Ronning D.R., Guynet C, Hickman AB, Ton-Hoang B, Chandler M, Dyda F. (2008). Mechanism of IS200/IS605 family DNA transposases: Activation and transposon-directed target site selection. Cell. 132, 208-220.

Bailey T.L. and Gribskov M. (1998). Combining evidence using p-values: application to sequence homology searches. Bioinformatics. 14, 48-54.

Berg D.E., Davies J., Allet B., Rochaix J.D. (1975) Transposition of R factor genes to bacteriophage λ. Proc. Natl. Acad. Sci. USA. 72, 3628-32.

Berkmen M.B. and Grossman A.D. (2006) Spatial and temporal organization of the *Bacillus subtilis* replication cycle. Mol Microbiol. 62, 57-71.

Besag J. and Clifford P. (1991). Sequential Monte Carlo p-values. Biometrika. 78, 301-304.

Beuzón C.R., Marqués S., Casadesús J. (1999). Repression of IS*200* transposase synthesis by RNA secondary structures. Nucleic Acids Res. 27, 3690-3695.

Bhasin A., Goryshin I.Y., Reznikoff W.S. (1999). Hairpin formation of Tn*5* transposition. J. Biol. Chem. 274, 37021-37029.

Bichsel M., Barbour A. D., Wagner A. (2010). The early phase of a bacterial insertion sequence infection. Theor. Popul. Biol. 78, 278-88.

Blakely G., May G., McCulloch R., Arciszewska L. K., Burke M., Lovett S. T., Sherratt D. J. (1993). Two related recombinases are required for site-specific recombination at dif and cer in E. coli K12. Cell. 75, 351-361.

Bujacz, G. Jaskólski M., Alexandratos J., Wlodawer A., Merkel G., Katz R.A. Skalka A.M. (1995). High-resolution structure of the catalytic domain of avian sarcoma virus integrase. J. Mol. Biol. 253, 333-346.

Bunting K.A., Roe S.M., Pearl, L.H. (2003). Structural basis for recruitment of translesion DNA polymerase Pol IV/DinB to the β -clamp. EMBO J. 22, 5883-5892.

Bowman G.D., O'Donnell M., Kuriyan J. (2004). Structural analysis of a eukaryotic sliding DNA clamp-clamp loader complex. Nature. 429, 724-730.

Braam L.M., Goryshin I.Y., Reznikoff W.S. (1999). A mechanism of Tn*5* inhibition: Carboxyl-terminal dimerization. J. Biol. Chem. 274, 86-92.

Brookfield J.F.Y. (2005). The ecology of the genome - mobile DNA elements and their hosts. Nat Rev Genet. 6, 128-136.

Bruck I. and O'Donnell M. (2001) The ring-type polymerase sliding clamp family. Genome Biol. 2, 3001.1-3001.3.

Casacuberta E. and González J. (2013). The impact of transposable elements in environmental adaptation. Mol. Ecol. 22, 1503-1517.

Cerveau N., Leclercq S., Leroy E., Bouchon D., Cordaux R. (2011). Short and long-term evolutionary dynamics of bacterial insertion sequences: insights from Wolbachia endosymbionts. Genome Biol. Evol. 3, 1175-1186.

Cerveau N., Leclercq S., Bouchon D., Cordaux, R. (2011b) in Evolutionary Biology - Concepts, Biodiversity, Macroevolution and Genome Evolution (ed. Pontarotti, P.) 291–312 Springer-Verlag, Berlin Heidelberg.

Chandler M., de la Cruz F., Dyda F., Hickman A.B., Moncalian G., Ton-Hoang B. (2013). Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. Nature Rev. Microbiol. 11, 525-538.

Chandler M. and Fayet O. (1993). Translational frameshifting in the control of transposition in bacteria. Mol. Microbiol. 7, 497-503.

Clemente J. C., Pehrsson E. C., *et al.* (2015). The microbiome of uncontacted Amerindians. Science Advances. 1(3).

Couturier E. and Rocha E.P.C. (2006). Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol. Microbiol. 59, 1506-1518.

Cordaux R., Udit S., Batzer M.A., Feschotte C. (2006). Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc. Natl. Acad. Sci. USA. 103, 8101-8106.

Craig N.L., Craigie R., Gellert M., Lambowitz A. (2002). Mobile DNA II. American Society of Microbiology, Washington, DC.

Curcio M. J. and Derbyshire K. M. (2003). The outs and ins of transposition: from Mu to Kangaroo. Nat. Rev. Mol. Cell. Biol. 4, 865-877.

Dalrymple B. P., Kongsuwan K., Wijffels G., Dixon N. E., Jennings P. (2001). A universal protein-protein interaction motif in the eubacterial DNA replication and repair systems. Proc. Natl. Acad. Sci. USA. 98, 11627-11632.

Davey M.J. and O'Donnell, M. Mechanisms of DNA replication. (2000). Curr. Opin. Chem. Biol. 4, 581–586.

Davies D.R., Braam L.M., Reznikoff W.S , Rayment I. (1999). The three-dimensional structure of a Tn*5* transposase-related protein determined to 2.9-Å resolution. J. Biol. Chem. 274, 11904-11913.

Davies D.R. Goryshin I.Y., Reznikoff W.S., Rayment I. (2000). Three-dimensional structure of the Tn5 synaptic complex transposition intermediate. Science. 289, 77-85.

Dawson A. and Finnegan D. J. (2003). Excision of the Drosophila mariner transposon Mos1: comparison with bacterial transposition and V(D)J recombination. Molecular Cell. 11, 225-235.

De la Cruz N.B., Weinreich M.D., Wiegand T.W., Krebs M.P., Reznikoff W.S. (1993). Characterization of the Tn5 transposase and inhibitor proteins: a model for the inhibition of transposition. J. Bacteriol. 175, 6932-38.

De Palmenaer D., Siguier P., Mahillon J. (2008) IS*4* family goes genomic. BMC Evol. Biol. 8, 18

Dobrindt U., Hochhut B., Hentschel U., Hacker J. (2004). Genomic islands in pathogenic and environmental microorganisms. Nat Rev. Micro. 2, 414-424.

Duval-Valentin G., Marty-Cointin B., Chandler M. (2004). Requirement of IS*911* replication before integration defines a new bacterial transposition pathway. EMBO J. 23, 3897-3906.

Dyda F., Hickman A.B., Jenkins T.M., Engelman A., Craigie R., Davies D.R. (1994). Crystal structure of the catalytic domain of HIV-1 integrase: Similarity to other polynucleotidyl transferases. Science. 266, 1981-1986.

Dyda F., Chandler M., Hickman A.B. (2012). The emerging diversity of transpososome architectures. Q Rev. Biophys. 45, 493-521.

Edgar R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 26, 2460-2461.

Escoubas J.M., Prere M.F., Fayet O., Salvignol I., D. Galas, Zerbib D., Chandler M. (1991). Translational control of transposition activity of the bacterial insertion sequence IS*1*. EMBO J. 10 705-712.

Fehér T., Bogos B., Méhi O., *et al.* (2012). Competition between transposable elements and mutator genes in bacteria. Mol. Biol. Evol. 29, 3153-3159.

Feil E.J. (2004). Small change: keeping pace with microevolution. Nat. Rev. Microbiol. 2, 483-495.

Filée J., Siguier P., Chandler M. (2007). Insertion sequence diversity in Archaea. Microbiol. Mol. Biol. Rev. 71, 121-157.

Garcillán-Barcia M.P., Bernales I., Mendiola M.V., de la Cruz F. (2001). Single-stranded DNA intermediates in IS*91* rolling-circle transposition. Mol Microbiol. 39, 494-501.

Garcillán-Barcia M.P. and de la Cruz F. (2002). Distribution of IS91 family insertion sequences in bacterial genomes: evolutionary implications. FEMS Microbiol. Ecol. 42, 303-313.

Gilbert C. and Cordaux R. (2013). Horizontal transfer and evolution of prokaryote transposable elements in eukaryotes. Genome Biol. Evol. 5, 822-832.

Georgescu R.E., Kim S.S., Yurieva O., Kuriyan J., Kong X.P., O´Donnell M. (2008). Structure of a sliding clamp on DNA. Cell. 132, 43-54.

Georgescu R. E., Yurieva O., Kim S.S, Kuriyan J., Kong X.P., O´Donnell M. (2008b). Structure of a small-molecule inhibitor of a DNA polymerase sliding clamp. Proc. Natl. Acad. Sci. USA. 105, 11116-11121.

Georgescu R., Langston L., O'Donnell M. (2015). A proposal: Evolution of PCNA's role as a marker of newly replicated DNA. DNA Repair. 29, 4-15.

Goryshin, I. Y. and Reznikoff W. S. (1998). Tn*5* in Vitro Transposition. J. Biol. Chem.. 273, 7367-7374.

Gouy M., Guindon S., Gascuel, O. (2010). SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27, 221-224.

Gradman R.J. and Reznikoff W.S. (2008). Tn5 synaptic complex formation: the role of transposase residue W450. J. Bacteriol. 190, 1484-87.

Griep M.A., McHenry C.S. (1988). The dimer of the beta subunit of *Escherichia coli* DNA polymerase III holoenzyme is dissociated into monomers upon binding magnesium(II). Biochemistry. 27, 5210-5215.

Gueguen E., Rousseau P., Duval-Valentin G., Chandler M. (2006). Truncated forms of IS911 transposase downregulate transposition. Mol Microbiol. 62, 1102-1116.

He S., Guynet C., Siguier P., Hickman A.B., Dyda F., Chandler M., Ton-Hoang B. (2013). IS200/IS605 family single-strand transposition: mechanism of IS608 strand transfer. Nucl Acids Res. 41, 3302-3313.

Halford S.E. and Marko J.F. (2004). How do site-specific DNA-binding proteins find their targets? Nucleic Acids Res. 32, 3040-52

Hedglin M., Kumar R., Benkovic S. J. (2013). Replication clamps and clamp loaders. Cold Spring Harb. Perspect. Biol. 5, a010165.

Hendrix R.W. (2003). Bacteriophage genomics. Curr. opin. microbiol. 506-11.

Hewson I., Poretsky R.S., Beinart R.A., White A.E. *et al.* (2009). In situ transcriptomic analysis of the globally important keystone N2-fixing taxon Crocosphaera watsonii. ISME J. 3, 618-631.

Hickman A.B., Chandler M., Dyda F. (2010). Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. Crit. Rev. Biochem. Mol. 45, 50-69.

Hickman A.B. and Dyda F. (2014). Mechanisms of DNA transposition. Microbiol Spectrum 3, 2. MDNA3-0034-2014.

Hiom K., Melek M., Gellert M. (1998). DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. Cell. 94, 463-470.

Hooper S. D., Mavromatis K. and Kyrpides, N. C. (2009). Microbial co-habitation and lateral gene transfer: what transposases can tell us. Genome Biol. 10, R45.

Hotopp J. C. D., M. E. Clark *et al.* (2007). Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science. 317, 1753-1756.

Hu W.Y. and Derbyshire K.M. (1998). Target choice and orientation preference of the insertion sequence IS*903*. J Bacteriol. 180, 3039-3048.

Indiani C. and M. O'Donnell (2006). The replication clamp-loading machine at work in the three domains of life. Nat. Rev. Mol. Cell Biol. 7, 51-761.

Iranzo J., Gómez M.J., López de Saro F.J., Manrubia S. (2014). Large-scale genomic analysis suggests a neutral punctuated dynamics of transposable elements in bacterial genomes. PLoS Comput Biol 10(6): e1003680.

Isberg R.R. and Syvanen M. (1982). DNA gyrase is a host factor required for transposition of Tn5. Cell. 30, 9-18.

Jang S., Sandler S.J., Harshey R.M. (2012). Mu insertions are repaired by the double-strand break repair pathway of Escherichia coli. PLoS Genet. 8, e1002642.

Jarvis T.C., Paul L.S., von Hippel P.H. (1989). Structural and enzymatic studies of the T4 DNA replication system. I. Physical characterization of the polymerase accessory protein complex. J Biol Chem. 264, 12709-12716.

Jeruzalmi D., Yurieva O., Zhao Y., Young M., Stewart J., Hingorani M., O'Donnell M., Kuriyan J. (2001a). Mechanism of processivity clamp opening by the delta subunit wrench of the clamp loader complex of *E. coli* DNA polymerase III. Cell. 106, 417-428.

Jeruzalmi D., O'Donnell M., Kuriyan J. (2001b). Crystal structure of the processivity clamp loader gamma complex of *E. coli* DNA polymerase III. Cell. 106, 429-441.

Johnson R.C. and Reznikoff W.S. (1983). DNA sequences at the ends of transposon Tn5 required for transposition. Nature. 304, 280-282.

Juhas M., van der Meer J.R., Gaillard M.,*et al.* (2009). Genomic islands, tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev. 33, 376-393.

Kennedy A.K., Guhathakurta A., Kleckner N., Haniford D.B. (1998). Tn*10* transposition via a DNA hairpin intermediate. Cell. 95, 125-134.

Kennedy A.K., Haniford D.B., Mizuuchi K. (2000). Single active site catalysis of the successive phosphoryl transfer steps by DNA transposases: Insights from phosphorothioate stereoselectivity. Cell. 101, 295-305.

Kichenaradja P., Siguier P., Perochon J., Chandler M. (2010). ISbrowser: an extension of ISfinder ulfor visualizing insertion sequences in prokaryotic genomes. Nucleic Acids Res. 38, D62–D68.

Kidwell M.G. and Lisch D. (1997). Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci U S A. 94, 7704-7711.

Kleckner N. (1989) in Mobile DNA (eds Berg, D. E. & Howe, M. M.) 227-268.

Kleckner N. (1990). Regulating Tn10 and IS10 transposition. Genetics. 124, 449-454.

Kong X.P., Onrust R., O'Donnell M., Kuriyan,J. (1992). Three dimensional structure of the β subunit of *Escherichia coli* DNA polymerase III Holoenzyme: a sliding DNA-clamp. Cell. 69, 425-437.

Kornberg A. and Baker T. A. (1992). DNA Replication. 2nd edn, Freeman,W. H., pp. 471-494. Springer, New York.

Krebs M.P. and Reznikoff W.S. (1986). Transcriptional and translational initiation sites of IS50. Control of transposase and inhibitor expression, J. Mol. Biol. 192, 781-791.

Krishna T.S., Kong X.P., Gary S., Burgers P.M., Kuriyan J. (1994). Crystal structure of the eukaryotic DNA polymerase processivity factor PCNA. Cell. 79, 1233-43.

Kubo K.M. and Craig, N.L. (1990). Bacterial transposon Tn7 utilizes two classes of target sites. J. Bacteriol. 172: 2774-2778.

Kunst F., N. Ogasawara, *et al.* (1997). The complete genome sequence of the Gram-positive bacterium Bacillus subtilis. Nature. 390, 249-256.

Le Rouzic A. and Capy P. (2005). The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. Genetics. 169, 1033-1043.

Levy S.B. and Marshall B. (2004). Antibacterial resistance worldwide: causes, challenges and responses. Nat. Med. 10. S122-129.

López de Saro F.J. and O'Donnell M. (2001). Interaction of the β sliding clamp with MutS, ligase, and DNA polymerase I. Proc. Natl. Acad. Sci. USA. 98, 8376-8380.

López de Saro F.J., Georgescu R.E., Goodman M.F., O'Donnell M. (2003). Competitive processivity-clamp usage by DNA polymerases during DNA replication and repair. EMBO J. 22, 6408-6418.

López de Saro F.J., Marinus M.G., Modrich P., O'Donnell M. (2006). The beta sliding clamp binds to multiple sites within MutL and MutS. J. Biol. Chem. 281, 14340-14349.

López de Saro F.J. (2009). Regulation of interactions with sliding clamps during DNA replication and repair. Curr Genomics. 10, 206-215.

López de Saro F.J., Gómez M., González-Tortuero E., Parro V. (2013). The dynamic genomes of acidophiles. In, J. Seckbach, A. Oren, H. Stan-Lotter, eds. Polyextremophiles, Life under multiple forms of stress. 1st ed. Dordrecht, Springer, 83-97.

López de Saro F.J., Díaz-Maldonado H., Amils R. (2015). Microbial evolution: the view from the acidophiles. In, Microbial Evolution under Extreme Conditions, Ed Corien Bakermans.

Lovell S., Goryshin I.Y., Reznikoff W.R., Rayment I. (2002). Two-metal active-site binding of a Tn5 transposase synaptic complex. Nature Struct. Biol. 9, 278-281.

Ma C. and Simons R.W. (1990). The IS10 antisense RNA blocks ribosome binding at the transposase translation initiation site, EMBO J. 9, 1267-1274.

Mahillon, J. and Chandler M.. (1998). Insertion sequences. Microbiol. Mol. Biol. R. 62, 725-774.

Malki M., De Lacey A. L., Rodríguez N., Amils R., Fernandez, V. M. (2008). Preferential use of an anode as an electron acceptor by an acidophilic bacterium in the presence of oxygen. App. Environ. Microbiol. 74, 4472-4476.

Martinez E. and de la Cruz F. (1990). Genetic elements involved in Tn*21* site-specific integration, a novel mechanism for the dissemination of antibiotic resistance genes. EMBO J. 9, 1275-1281.

Martusewitsch E., Sensen C.W., Schleper C. (2000). High spontaneous mutation rate in the hyperthermophilic archaeon Sulfolobus solfataricus is mediated by transposable elements. J Bacteriol. 182, 2574-81.

Mátés L., Chuah M.K.L., Belay E. *et al.* (2009). Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. Nat. Genet. 41, 753-761.

Matsumiya S, Ishino Y, Morikawa K. (2001). Crystal structure of an archaeal DNA sliding clamp: Proliferating cell nuclear antigen from Pyrococcus furiosus. Protein Sci. 10, 17-23.

Maul R.W., Ponticelli S.K., Duzen, J.M., Sutton M.D. (2007). Differential binding of *Escherichia coli* DNA polymerases to the beta sliding clamp. Mol. Microbiol. 65, 811-827.

McClintock B. (1950). The origin and behavior of mutable loci in maize. Proc.Natl. Acad. Sci. USA. 36, 344-355.

McInerney P., Johnson A., Katz F., O'Donnell, M. (2007). Characterization of a triple DNA polymerase replisome. Mol.Cell. 27, 527-538.

Medini D., Donati C., Tettelin H., Masignani V., Rappuoli R. (2005). The microbial pan-genome. Curr Opin Genet Dev. 15, 589-94.

Migocki M.D., Lewis P.J., Wake R.G., Harry E.J. (2004).The midcell replication factory in *Bacillus subtilis* is highly mobile: implications for coordinating chromosome replication with other cell cycle events. Mol. Microbiol. 54, 452-463.

Mizuuchi K. and Adzuma K. (1991). Inversion of the phosphate chirality at the target site of Mu DNA strand transfer: evidence for a one-step transesterification mechanism. Cell. 66, 129-140.

Mizuuchi K. and Baker T.A. (2002). Chemical mechanisms for mobilizing DNA. In: Craig NL, Craigie R, Gellert M and Lambowitz AM, eds, Mobile DNA II, pp. 12-23. Washington, DC, ASM Press.

Moarefi I., Jeruzalmi D., Turner J., O'Donnell M., Kuriyan J. (2000). Crystal structure of the DNA polymerase processivity factor of T4 bacteriophage. J Mol Biol. 296, 1215-1223.

Moldovan G.L., Pfander B., Jentsch S. (2007). PCNA, the maestro of the replication fork. Cell. 129, 665-679.

Mongodin, E.F., Nelson K.E., *et al.* (2005). The genome of Salinibacter ruber: convergence and gene exchange among hyperhalophilic bacteria and archaea. Proc.Natl. Acad. Sci. USA. 102, 18147-18152.

Montaño S.P., Pigli Y.Z., Rice P.A. (2012). The Mu transpososome structure sheds light on DDE recombinase evolution. Nature. 491, 413-417.

Moolman M.C., Krishnan S.T., Kerssemakers J.W.J. *et al.* (2014). Slow unloading leads to DNA-bound β2-sliding clamp accumulation in live *Escherichia coli* cells. Nat Commun 5.2014

Moreno-Paz M. and Parro V. (2006). Amplification of low quantity bacterial RNA for microarray studies: time-course analysis of *Leptospirillum ferrooxidans* under nitrogenfixing conditions. Environ Microbiol. 8, 1064-1073.

Mott M.L and Berger J.M. (2007). DNA replication initiation: mechanisms and regulation in bacteria. Nature Reviews Microbiology. 5, 343-354.

Mueller R. S., Denef V.J., Kalnejais L.H., Suttle K.B. *et al.* (2010). Ecological distribution and population physiology defined by proteomics in a natural microbial community. Mol. Syst. Biol. 6, 374.

Nagy Z. and Chandler M. (2004). Regulation of transposition in bacteria. Res. Microbiol. 155, 387-398.

Nakai H., Doseeva V., Jones J.M. (2001). Handoff from recombinase to replisome: insights from transposition. Proc Natl Acad Sci U S A. 98, 8247-8254.

Nelson D. and Cox M. (2004). Lehninger principles of biochemistry (4th ed.). M. Freeman, ed.

Nicolas E., Lambin M., Dandoy D., Galloy C., Nguyen N., Oger C.A., Hallet B. (2015). The Tn3-family of replicative transposons. Microbiology Spectrum. 3, 2.

Oliver K.R. and Green W.K. (2009). Transposable elements: powerful facilitators of evolution. BioEssays. 31, 703-714.

Onrust R., Stukenberg P.T., O´Donnell M. (1991). Analysis of the ATPase subassembly which initiates processive DNA synthesis by DNA polymerase III holoenzyme. J. Biol. Chem. 266, 21681-21686.

Onrust R., Finkelstein J., Naktinis V., Turner J., Fang L. O´Donnell M. (1995). Assembly of a chromosomal replication machine: two DNA polymerases, a clamp loader and sliding clamps in one holoenzyme particle. I. Organization of the clamp loader. J. Biol. Chem. 270, 13348-13357.

Papke R.T., Koenig J.E., Rodríguez-Valera F., Doolittle, W.F. (2004). Frequent recombination in a saltern population of Halorubrum. Science. 306, 1928-1929.

Parks A.R., Li Z., Shi Q., Owens R.M., Jin M.M., Peters J.E. (2009). Transposition into replicating DNA occurs through interaction with the processivity factor. Cell. 138, 685-695.

Parro V., Moreno-Paz M., González-Toril E. (2007). Analysis of environmental transcriptomes by DNA microarrays. Environ Microbiol. 9, 453-464.

Partridge S.R., Hall R.M. (2003). The IS1111 family members IS4321 and IS5075 have subterminal inverted repeats and target the terminal inverted repeats of Tn21 family transposons. J. Bacteriol. 185, 6371-6384.

Paul S., Million-Weaver S., Chattopadhyay S., Sokurenko E., Merrikh H. (2013). Accelerated gene evolution through replication-transcription conflicts. Nature. 495, 512-515.

Peña A., Teeling H., Huerta-Cepas J. *et al.* (2010). Fine-scale evolution: genomic, phenotypic and ecological differentiation in two coexisting Salinibacter ruber strains. ISME J. 4, 882-895.

Peters J. and Craig N.L. (2001). Tn*7* recognizes transposition target structures associated with DNA replication using DNA binding protein TnsE. Genes Dev., 15, 737-747.

Pisani F.M., De Felice M., Carpentieri F., Rossi M. (2000). Biochemical characterization of a clamp-loader complex homologous to eukaryotic replication factor C from the hyperthermophilic archaeon Sulfolobus solfataricus. J. Mol. Biol. 301, 61–73.

Plague G.R. (2010). Intergenic transposable elements are not randomly distributed in bacteria. Genome Biol Evol. 2, 584-590.

Post V. and Hall R.M. (2009). Insertion sequences in the IS1111 family that target the attC recombination sites of integron-associated gene cassettes. FEMS Microbiol Lett. 290, 182-187

Pritchard A.E., Dallmann H.G., Glover B.P., McHenry C.S. (2000). A novel assembly mechanism for the DNA polymerase III holoenzyme DnaX complex: association of δδ´ with DnaX(4) forms DnaX(3) δδ´. EMBO J. 19, 6536-6545.

Punta M., Coggill P.C., Eberhardt R.Y., et al. (2012). The Pfam protein families database. Nucleic Acids Res. 40. D290-D301.

Ram R.J., VerBerkmoes N.C., Thelen M.P., Tyson G.W., Baker B.J., *et al.* (2005). Community proteomics of a natural microbial biofilm. Science. 308, 1915-1920.

Reznikoff W.S. (1993) The Tn*5* transposon. Annu. Rev. Microbiol. 47, 945-63.

Reznikoff W.S. (2002). Tn*5* transposition. In Mobile DNA II. Ed. American Society of Microbiology. 403-421.

Reznikoff W.S., Bordenstein S.R., Apodaca J. (2004). Comparative Sequence Analysis of IS50/Tn5 Transposase. J. Bacteriol. 186, 8240-8247.

Reznikoff W.S., Steiniger-White M.M., Metzler J.D. (2006). Tn*5* transposase mutants and the use thereof. US Patent US7083980.

Reznikoff W.S. (2008). Transposon Tn*5*.Annu. Rev. Genet. 42, 269-286.

Rice P., Longden I., Bleasby A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16, 276-277.

Roberts D., Hoopes B.C., McClure W.R., Kleckner N. (1985). IS*10* transposition is regulated by DNA adenine methylation. Cell. 43,117-130.

Rocha E. (2002). Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends Microbiol. 10, 393-395.

Rocha E.P.C. and Danchin A. (2003). Essentiality, not expressiveness, drives gene strand bias in bacteria. Nat Genet. 34, 377-378.

Rocha E.P.C. (2003). DNA repeats lead to the accelerated loss of gene order in Bacteria. Trends Genet. 19, 600-604.

Rocha E.P.C., Touchon M., Feil E.J. (2006). Similar compositional biases are caused by very different mutational effects. Genome Res. 16, 1537-1547.

Rocha E.P.C. (2008). The organization of the bacterial genome. Annu. Rev. Genet. 42, 211-233.

Rogers M., Ekaterinaki N., Nimmo E., Sherratt D. (1986). Analysis of Tn*7* transposition. Mol. Gen. Genet. 205, 550-556.

Ronning D.R., Guynet C., Ton-Hoang B., Perez Z.N., Ghirlando R., Chandler M., Dyda F. (2005). Active site sharing and subterminal hairpin recognition in a new class of DNA transposases. Mol. Cell. 20, 143-154.

Rothstein S.J. and Reznikoff W.S. (1981). The functional differences in the inverted repeats of Tn*5* are caused by a single base pair nonhomology. Cell.23, 191-199.

Rutherford K., Parkhill J., Crook J. *et al.* (2000). Artemis: sequence visualization and annotation. Bioinformatics. 16, 944-945.

Sambrook J. and Russel D.W. (2001). Molecular Cloning: A Laboratory Manual. 3rd ed, Cold Spring Harbor Laboratory Press, New York.

San Martín-Uriz P., Gómez M.J., Arcas A., Bargiela R., Amils R. (2011). Draft genome sequence of the electricigen *Acidiphilium sp.* strain PM (DSM 24941). J. Bacteriol. 193, 5585-5586

Sanders G.M., Dallmann G., McHenry C.S. (2010). Reconstitution of the B. subtilis replisome with 13 proteins including two distinct replicases. Mol Cell. 37, 273-281.

Sasakawa C., Uno Y., Yoshikawa M. (1981). The requirement for both DNA polymerase and 5´ to 3´ exonuclease activities of DNA polymerase I during Tn5 transposition, Mol. Gen. Genet. 182, 19-24.

Schlor S., Reidl S., Blass J., Reidl J. (2000). Genetic arrangements of the regions adjacent to genes encoding heat-labile enterotoxins (eltAB) of enterotoxigenic Escherichia coli strains. Appl. Environ. Microbiol. 66, 352-358.

Schneider C.A., Rasband W.S., Eliceiri K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nature Methods 9, 671-675.

Schönknecht G., Chen W.H., Ternes C.M., Barbier G.G., *et al.* (2013). Gene transfer from bacteria and archaea facilitated evolution of an extremophilic Eukaryote. Science. 339, 1207-1210.

Schulz V. and Reznikoff W.S. (1991). Translation initiation of IS*50*R read-through transcripts. J. Mol. Biol. 221, 65-80.

Sekine Y., Aihara K., Ohtsubo E. (1999). Linearization and transposition of circular molecules of insertion sequence IS3. J Mol Biol. 294, 21-34

Skelding Z., Queen-Baker J., Craig N.L. (2003). Alternative interactions between the Tn*7* transposase and the Tn*7* target DNA binding protein regulate target immunity and transposition. EMBO J. 22, 5904-5917.

Siguier P., Perochon J., Lestrade L., Mahillon J., Chandler M. (2006). ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 34, D32-D36.

Siguier P, Gourbeyre E, Chandler M. (2014). Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol Rev. 38, 865-891.

Sobetzko P., Travers A., Muskhelishvili G. (2012). Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. Proc. Natl. Acad. Sci. USA. 109, E42-E50.

Sousa A., Bourgard C., Wahl L. M., Gordo I. (2013). Rates of transposition in *Escherichia coli*. Biol Lett. 9, 20130838.

Srivatsan A., Tehranchi A., MacAlpine D.M., Wang J.D. (2010). Co-orientation of replication and transcription preserves genome integrity. PLoS Genet. 6(1):e1000810.

Steiniger M., Adams C.D., Marko J.F., Reznikoff W.S. (2006). Defining characteristics of Tn5 transposase nonspecific DNA binding. Nucleic Acids Res. 34, 2820-32.

Steiniger M., Metzler J., Reznikoff W.S. (2006). Mutation of Tn5 transposase β-loop residues affects all steps of Tn5 transpostion: the role of conformational changes in Tn5 transposition. Biochemistry. 45, 15552-62.

Steiniger-White M. and Reznikoff W.S. (2000). The C-terminal alpha helix of Tn5 transposase is required for synaptic complex formation. J. Biol. Chem. 275, 23127-33.

Sternglanz R., DiNardo S., Voelkel K.A., Nishimura Y., Hirota Y., Becherer K., Zumstein L.,Wang J.C. (1981). Mutations in the gene coding for *Escherichia coli* DNA topoisomerase I affect transcription and transposition, Proc. Natl. Acad. Sci. USA. 78, 2747-2751.

Su'etsugu M. and Errington J. (2011). The replicase sliding clamp dynamically accumulates behind progressing replication forks in *Bacillus subtilis* cells. Mol Cell. 41, 720-732

Taylor M.S., LaCava J., Mita P. *et al.* (2013). Affinity proteomics reveals human host factors implicated in discrete stages of LINE-1 retrotransposition. Cell. 155, 1034-1048.

Tetu S.G. and Holmes A.J. (2008). A family of insertion sequences that impacts integrons by specific targeting of gene cassette recombination sites, the IS1111-attC group. J Bacteriol. 190, 4959-4970

Thomas C.M. and Nielsen,K.M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nature Rev. Microbiol. 3, 711-21.

Tobes R. and Pareja E. (2006). Bacterial repetitive extragenic palindromic sequences are DNA targets for insertion sequence elements. BMC Genomics. 7, 62.

Ton-Hoang B, Pasternak C, Siguier P, Guynet C, Hickman AB, Dyda F, Sommer S, Chandler M. Single-stranded DNA transposition is coupled to host replication. Cell. 142, 398–408. 2010

Touchon M. and Rocha E.P.C. (2007). Causes of insertion sequences abundance in prokaryotic genomes. Mol Biol Evol. 24, 969-981.

Touchon M., Hoede C., Tenaillon O., Barbe V., *et al.* (2009)Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. 5: e1000344.

Toussaint A. and Chandler M. (2012).Prokaryote genome fluidity, toward a system approach of the mobilome. Methods in molecular biology. 804, 57-80.

Turlan C., Loot C., Chandler M. (2004). IS*911* partial transposition products and their processing by the *Escherichia coli* RecG helicase. Mol Microbiol. 53, 1021-1033.

van Gent DC, Mizuuchi K, Gellert M. Similarities between initiation of V(D)J recombination and retroviral integration. Science 271, 1592–1594. 1996

van Hoek A.H., Mevius D., Guerra B., Mullany P., Roberts A.P., Aarts H.J.M. (2011). Acquired Antibiotic Resistance Genes: An Overview. Front Microbiol. 2, 203.

Van Valen L. (1974) Molecular evolution as predicted by natural selection. J. Mol. Evol. 3, 89-101.

Venkatesan M.M., Goldberg M.B., Rose D.J., Grotbeck E.J., Burland V., Blattner F.R. (2001). Complete DNA sequence and analysis of the large virulence plasmid of Shigella flexneri. Infect Immun. 69, 3271-3285.

Waddell C.S. and Craig N.L. (1988). Tn7 transposition: two transposition pathways directed by five Tn7-encoded genes. Genes Dev. 2, 137-149.

Wagner A. (2006). Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. Mol. Biol. Evol. 23, 723-733.

Wagner A. and de la Chaux N. (2008). Distant horizontal gene transfer is rare for multiple families of prokaryotic insertion sequences. Mol. Genet. Genomics. 280, 397-408.

Wagner A. (2009). Transposable elements as genomic diseases. Mol. BioSyst. 5, 32-35.

Wang J.D., Berkmen M.B., Grossman A.D. (2007). Genome-wide co-orientation of replication and transcription reduces adverse effects on replication in Bacillus subtilis. Proc Natl Acad Sci USA. 104, 5608-5613.

Warbrick E., Heatherington, W., Lane, D. P., Glover, D. M. (1998). PCNA binding proteins in Drosophila melanogaster: the analysis of a conserved PCNA binding domain. Nuc. Acids Res. 26, 3925-32.

Warbrick E. (2000). The puzzle of PCNA's many partners. Bioessays. 22, 997-1006.

Weinreich M.D., Yigit H., Reznikoff W.S. (1994a). Overexpression of the Tn5 Transposase in Escherichia coli Results in Filamentation, Aberrant Nucleoid Segregation, and Cell Death: Analysis of E. coli and Transposase Suppressor Mutations. J Bacteriol. 176, 5494-5504.

Weinreich M.D., Mahnke-Braam L, Reznikoff W.S. (1994b). A functional analysis of the Tn5 transposase: identification of domains required for DNA binding and multimerization. J. Mol. Biol. 241, 166-77.

Weinreich M.D., Gasch A, Reznikoff WS. (1994c). Evidence that the cis preference of the Tn5 transposase is caused by nonproductive multimerization. Genes Dev. 8, 2363-74.

Wicker T., Sabot F., Hua-van A., Bennetzen J.L. *et al.* (2007). A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8, 973-982.

Wiegand T.W. and Reznikoff W.S. (1992). Characterization of two hypertransposing Tn5 mutants. J. Bacteriol. 174, 1229-1239.

Williams D., Gogarten J.P., Papke R.T. (2012). Quantifying homologous replacement of loci between haloarchaeal species. Genome Biol. Evol. 4, 1223-1244.

Wolkow C.A., DeBoy R.T., Craig N.L. (1996). Conjugating plasmids are preferred targets for Tn7. Genes Dev. 10, 2145-2157.

World Health Organization. (2014). Antimicrobial resistance: global report on surveillance.

Wozniak R.A. and Waldor M.K. (2010). Integrative and conjugative elements, mosaic mobile genetic elements enabling dynamic lateral gene flow. Nat Rev Microbiol, 8, 552-563.

Yao N., Turner J., Kelman Z. *et al.* (1996). Clamp loading, unloading and intrinsic stability of the PCNA, β and gp45 sliding clamps of human, *E. coli* and T4 replicases. Genes Cells. 1, 101-113.

York D. and Reznikoff W.S. (1996). Purification and biochemical analysis of a monomeric form of Tn5 transposase. Nucleic Acids Res. 24, 3790-96.

Yin J.C., Krebs M.P., Reznikoff W.S. (1988). Effect of dam methylation on Tn5 transposition. J Mol Biol. 199, 35-45.

Yusa K., Zhou L., Li M.A., Bradley A., Craig N.L. (2011). A hyperactive piggyBac transposase for mammalian applications. Proc. Natl. Acad. Sci. USA. 108, 1531-1536.

Zerbib D., Gamas P., Chandler M. *et al.* (1985). Specificity of insertion of IS*1*. J Mol Biol. 185, 517-524.

Zhou M. and Reznikoff W.S. (1997). Tn*5* transposase mutants that alter DNA binding specificity. J. Mol. Biol. 271,362-73.

Zhou M., Bhasin A., Reznikoff, W.S. (1998). Molecular genetic analysis of transposase-end DNA sequence recognition: cooperativity of three adjacent base-pairs in specific interaction with a mutant Tn*5* transposase. J. Mol. Biol. 276, 913-925.

Appendices

8.1. Appendix I

Table I.1. List of Pfam domains used to identify bacterial transposases

DDE_2	DUF4158	IstB_IS21_ATP
DDE_3	DUF4277	LZ_Tnp_IS481
DDE_4	DUF4338	LZ_Tnp_IS66
DDE_4_2	DUF4351	MULE
DDE_5	DUF4372	Mu-transpos_C
DDE_Tnp_1	DUF772	Nterm_IS4
DDE_Tnp_1_2	HTH_17	OrfB_IS605
DDE_Tnp_1_3	HTH_21	OrfB_Zn_ribbon
DDE_Tnp_1_4	HTH_23	Phage-MuB_C
DDE_Tnp_1_5	HTH_24	Resolvase
DDE_Tnp_1_6	HTH_28	rve
DDE_Tnp_1_assoc	HTH_29	rve_2
DDE_Tnp_2	HTH_32	rve_3
DDE_Tnp_IS1	HTH_33	Tn7_Tnp_TnsA_C
DDE_Tnp_IS1595	HTH_38	Tn7_Tnp_TnsA_N
DDE_Tnp_IS240	HTH_7	Tn7_TnsC_Int
DDE_Tnp_IS66	HTH_OrfB_IS605	TnpB_IS66
DDE_Tnp_IS66_C	HTH_Tnp_1	Transposase_20
DDE_Tnp_ISAZ013	HTH_Tnp_IS1	Transposase_31
DDE_Tnp_ISL3	HTH_Tnp_IS630	Transposase_mut
DDE_Tnp_Tn3	HTH_Tnp_IS66	Y1_Tnp
DEDD_Tnp_IS110	HTH_Tnp_ISL3	Y2_Tnp
Dimer_Tnp_Tn5	HTH_Tnp_Mu_1	Zn_Tnp_IS1
DUF2080	HTH_Tnp_Mu_2	Zn_Tnp_IS1595
DUF4096	IstB_IS21	Zn_Tnp_IS91

Table I.2. Pfam-based architecture description of IS structures and classification into IS families.

The table lists 209 gene clusters corresponding to IS elements. Gene symbols are enclosed by characters that represent an arrow (--gene->), to describe their relative orientation, and are connected by double colons (::). The gene symbols represent the domain organization of encoded proteins. Domain names follow Pfam nomenclature and are connected by equal signs (=). IS family nomenclature follows that of IS finder (www-is.biotoul.fr//is.html), except for IS*4*, IS*5*, IS*66*, and IS*NCY*, as described in accompanying Notes. Importantly, our classification aims to group the domain architectures and proteins of each IS family according to the transposase domains detected and does not take into account other aspects of IS structure, such as the flanking inverted repeats present in many ISs.

IS FAMILY	IS STRUCTURE	NOTES
IS*1*	--DDE_Tnp_IS1-> --Zn_Tnp_IS1=HTH_Tnp_IS1->::--DDE_Tnp_IS1-> --Zn_Tnp_IS1=HTH_Tnp_IS1-> --Zn_Tnp_IS1=DDE_Tnp_IS1-> --HTH_23=DDE_Tnp_IS1-> --Zn_Tnp_IS1=HTH_Tnp_IS1=DDE_Tnp_IS1-> --HTH_Tnp_IS1->::--DDE_Tnp_IS1-> --Zn_Tnp_IS1->::--DDE_Tnp_IS1->	
IS*3*	--HTH_Tnp_1->::--HTH_21=rve-> --HTH_21=rve-> --HTH_Tnp_1->::--HTH_21=rve_3-> --HTH_29=rve-> --rve_3-> --HTH_28=HTH_28->::--HTH_21=rve-> --rve_2-> --HTH_Tnp_1=HTH_21=rve-> --HTH_28->::--HTH_21=rve-> --HTH_Tnp_1->::--rve_3-> --HTH_21=rve_3-> --HTH_Tnp_1->::--HTH_21-> --HTH_28=HTH_21=rve-> --HTH_28=rve-> --HTH_28->::--HTH_21=rve_2-> --HTH_21=rve_2-> --HTH_32=rve-> --HTH_28=rve_3-> --HTH_Tnp_1=HTH_28->::--HTH_21=rve-> --HTH_Tnp_1=HTH_21=rve_3-> --HTH_Tnp_1->::--HTH_29=rve-> --HTH_29=rve_3-> --HTH_28=HTH_Tnp_1=HTH_28->::--HTH_21=rve-> --HTH_21->::--rve_2-> --HTH_29=rve=rve_2-> --HTH_Tnp_1->::--HTH_21=rve_2-> --HTH_Tnp_1=rve-> --HTH_32=rve_3-> --HTH_Tnp_1->::--HTH_21->::--rve_3-> --HTH_28=HTH_21=rve_2-> --HTH_38=HTH_Tnp_1->::--HTH_21=rve_2-> --HTH_28->::--rve_2-> --HTH_28=HTH_28=HTH_28->::--HTH_21=rve-> --HTH_21->::--rve_3-> --HTH_Tnp_1->::--HTH_32=rve->	

	<-HTH_Tnp_1--::--rve_3-> --HTH_28=HTH_28=HTH_21=rve-> --HTH_28->::--HTH_21->::--rve_2-> --HTH_28->::--rve_3-> --HTH_Tnp_1->::--HTH_38=rve-> --HTH_Tnp_1->::<-HTH_38=rve-- --HTH_29=rve=HTH_29-> --HTH_Tnp_1=rve_3-> --HTH_23=HTH_38=rve-> --HTH_28=HTH_28->::--HTH_21->::--rve_2-> --HTH_Tnp_1->::--rve_2-> --HTH_23=HTH_21=rve->	
IS*4*a	--Nterm_IS4=DDE_Tnp_1-> --Nterm_IS4-> --Nterm_IS4->::--DDE_Tnp_1->	Includes members of subgroup IS*4* of family IS*4*
IS*4*b	--DUF4372=DDE_Tnp_1-> --DUF4372-> --DUF4372->::--DDE_Tnp_1->	Includes members of subgroup IS*4Sa* of family IS*4*
IS*4*c	--Dimer_Tnp_Tn5-> --DDE_Tnp_1=Dimer_Tnp_Tn5->	Includes members of the IS*50* subgroup of family IS*4*
IS*5*a	--DUF772=DDE_Tnp_1-> --DUF772-> --DUF772=DDE_Tnp_1_2-> --DUF772->::--DDE_Tnp_1-> --DUF772->::--DDE_Tnp_1_2->	Includes members of the IS*5* subgroup of family IS*5*
IS*5*b	--DUF4096->::--DDE_Tnp_1_2-> --DUF4096-> --DUF4096=DDE_Tnp_1-> --DUF4096=DDE_Tnp_1_2-> --DUF4096->::--DDE_Tnp_1->	Includes members of subgroups IS*427* and IS*1031* of family IS*5*
IS*5*c	--DDE_Tnp_1_5->	Includes members of subgroup ISL2 of family IS5
IS*5*d	--DDE_4-> --DDE_4_2-> --DDE_4_2->::--DDE_4-> --DDE_4_2=DDE_4->	Includes members of subgroup IS903 of family IS5
IS*6*	--DDE_Tnp_IS240-> --Zn_Tnp_IS1=DDE_Tnp_IS240->	
IS*21*	--IstB_IS21-> --HTH_38=rve->::--IstB_IS21-> --IstB_IS21_ATP=IstB_IS21-> --HTH_23=rve->::--IstB_IS21-> --HTH_7=rve->::--IstB_IS21-> --HTH_23->::--IstB_IS21-> --HTH_23=rve->::--IstB_IS21_ATP=IstB_IS21-> --HTH_29=rve->::--IstB_IS21-> --IstB_IS21->::--HTH_21=rve->	
IS*30*	--HTH_38=rve-> --HTH_38->	
IS*66*a	--HTH_Tnp_1->::--TnpB_IS66->::-- LZ_Tnp_IS66=HTH_Tnp_IS66=DDE_Tnp_IS66=DDE_Tnp_IS66_C-> --TnpB_IS66->::-- LZ_Tnp_IS66=HTH_Tnp_IS66=DDE_Tnp_IS66=DDE_Tnp_IS66_C-> --TnpB_IS66->::--LZ_Tnp_IS66=HTH_Tnp_IS66=DDE_Tnp_IS66->	

	--LZ_Tnp_IS66=HTH_Tnp_IS66=DDE_Tnp_IS66=DDE_Tnp_IS66_C-> --LZ_Tnp_IS66=HTH_Tnp_IS66=DDE_Tnp_IS66-> --TnpB_IS66->::--LZ_Tnp_IS66=DDE_Tnp_IS66=DDE_Tnp_IS66_C-> --HTH_Tnp_1->::--TnpB_IS66->::-- LZ_Tnp_IS66=HTH_Tnp_IS66=DDE_Tnp_IS66-> --HTH_Tnp_1->::-- LZ_Tnp_IS66=HTH_Tnp_IS66=DDE_Tnp_IS66=DDE_Tnp_IS66_C-> --TnpB_IS66->::--LZ_Tnp_IS66->::-- HTH_Tnp_IS66=DDE_Tnp_IS66=DDE_Tnp_IS66_C-> --TnpB_IS66->::--LZ_Tnp_IS66->::--HTH_Tnp_IS66=DDE_Tnp_IS66-> --TnpB_IS66->::--HTH_Tnp_IS66=DDE_Tnp_IS66-> --TnpB_IS66->::--LZ_Tnp_IS66=DDE_Tnp_IS66-> --HTH_Tnp_1->::--TnpB_IS66->::--LZ_Tnp_IS66-> --HTH_Tnp_IS66=DDE_Tnp_IS66=DDE_Tnp_IS66_C-> --TnpB_IS66->::--HTH_Tnp_IS66=DDE_Tnp_IS66=DDE_Tnp_IS66_C-> --HTH_23=HTH_Tnp_1->::--TnpB_IS66->::-- LZ_Tnp_IS66=HTH_Tnp_IS66=DDE_Tnp_IS66=DDE_Tnp_IS66_C-> --TnpB_IS66->::--LZ_Tnp_IS66=HTH_Tnp_IS66->	The IS*66* family was divided in two groups according to Gourbeyre *et al.*, 2010
IS*66*b	--HTH_Tnp_IS66=DDE_Tnp_IS66->	Includes the subgroup ISBst12 of IS66
IS*91*	--Zn_Tnp_IS91=Y2_Tnp-> --Y2_Tnp-> --Zn_Tnp_IS91->::--Y2_Tnp->	
IS*110*	--DEDD_Tnp_IS110=Transposase_20-> --Transposase_20-> --DEDD_Tnp_IS110-> --DEDD_Tnp_IS110->::--Transposase_20->	
IS*200*	--Y1_Tnp-> --Y1_Tnp=Y1_Tnp->	
IS*200*/IS*605*	<-HTH_OrfB_IS605=OrfB_IS605=OrfB_Zn_ribbon--::--Y1_Tnp-> --Y1_Tnp->::--HTH_OrfB_IS605=OrfB_IS605=OrfB_Zn_ribbon-> --Y1_Tnp->::--OrfB_IS605=OrfB_Zn_ribbon-> <-Y1_Tnp--::--OrfB_IS605=OrfB_Zn_ribbon-> <-OrfB_IS605--::--Y1_Tnp-> <-HTH_OrfB_IS605=OrfB_IS605--::--Y1_Tnp-> <-Y1_Tnp--::--HTH_OrfB_IS605->::--OrfB_IS605=OrfB_Zn_ribbon->	We have chosen TnpA (domain Y1_Tnp) as a proxy to score the orientation of the insertion sequence in this family
IS*256*	--Transposase_mut->	
IS*481*	--LZ_Tnp_IS481=rve-> --LZ_Tnp_IS481->	
IS*607*	--Resolvase->::--HTH_OrfB_IS605=OrfB_IS605=OrfB_Zn_ribbon-> --HTH_17=Resolvase->::--OrfB_IS605=OrfB_Zn_ribbon-> --MerR=Resolvase-> --Resolvase->::--HTH_OrfB_IS605=OrfB_IS605-> --HTH_17=Resolvase-> --MerR=Resolvase->::--HTH_OrfB_IS605=OrfB_IS605=OrfB_Zn_ribbon-> --MerR_1=Resolvase->::--OrfB_IS605=OrfB_Zn_ribbon-> --HTH_17=Resolvase->::--OrfB_Zn_ribbon-> --MerR=Resolvase->::--OrfB_Zn_ribbon-> --MerR_1=Resolvase->::--OrfB_IS605-> --Resolvase->::--OrfB_IS605=OrfB_Zn_ribbon-> --MerR_1=Resolvase->::-- HTH_OrfB_IS605=OrfB_IS605=OrfB_Zn_ribbon-> --MerR=Resolvase->::--OrfB_IS605=OrfB_Zn_ribbon-> --Resolvase->::--OrfB_IS605-> --HTH_17=Resolvase->::--OrfB_IS605->	
IS*630*	--HTH_29=DDE_3-> --DDE_3->	

	--HTH_Tnp_IS630=DDE_3-> --HTH_Tnp_IS630->::--DDE_3-> --HTH_29->::--DDE_3-> --HTH_23=HTH_33=DDE_3-> --HTH_Tnp_IS630-> --HTH_23=HTH_33->::--DDE_3-> --HTH_33=DDE_3-> --HTH_28->::--DDE_3-> --HTH_28=HTH_33->::--DDE_3-> --HTH_33->::--DDE_3-> --HTH_23->::--DDE_3-> --HTH_28=DDE_3-> --HTH_23=DDE_3-> --HTH_23->::--HTH_33->::--DDE_3->	
IS*701*	--DDE_5-> --DDE_5=DDE_Tnp_1-> --DDE_Tnp_1->::--DDE_5-> --DDE_5->::--DDE_Tnp_1-> --DDE_Tnp_1->::<-DDE_5-- <-DDE_Tnp_1--::--DDE_5->	
IS*982*	--DDE_Tnp_1_3-> --DDE_Tnp_1_3=DDE_Tnp_1_3->	
IS*1182*	--DUF772=DDE_Tnp_1_6-> --DDE_Tnp_1_6-> --DUF772->::--DDE_Tnp_1_6->	
IS*1380*	--DDE_Tnp_1_4->	
IS*1595*	--DDE_Tnp_IS1595-> --Zn_Tnp_IS1595=DDE_Tnp_IS1595-> --Zn_Tnp_IS1595->::--DDE_Tnp_IS1595->	
IS*1634*	--DUF4277->	
IS*As1*	--DUF4277=DDE_Tnp_1-> --DDE_Tnp_1_assoc=DDE_Tnp_1-> --DDE_Tnp_1_assoc-> --DDE_Tnp_1_assoc->::--DDE_Tnp_1-> --DUF4338=DDE_Tnp_1_assoc=DDE_Tnp_1-> --DUF4338->::--DDE_Tnp_1_assoc->	
IS*Azo13*	--DDE_Tnp_ISAZ013->	
IS*L3*	--HTH_Tnp_ISL3=DDE_Tnp_ISL3-> --DDE_Tnp_ISL3-> --HTH_Tnp_ISL3->::--DDE_Tnp_ISL3-> --HTH_Tnp_ISL3=DDE_Tnp_ISL3=DDE_Tnp_ISL3-> --DDE_Tnp_ISL3=DDE_Tnp_ISL3-> --HTH_Tnp_ISL3->::--DDE_Tnp_ISL3=DDE_Tnp_ISL3->	
IS*NCYa*	--Transposase_31-> --DUF4351-> --Transposase_31=DUF4351->	ISPlu15
IS*NCYb*	--MULE->	ISM1
IS*NCYc*	--DDE_Tnp_2->	ISC1217
IS*NCYd*	--DUF2080->	ISA1214
Tn*3*	--DUF4158=DDE_Tnp_Tn3-> --DDE_Tnp_Tn3-> --Resolvase=HTH_7->::--DUF4158=DDE_Tnp_Tn3-> <-Resolvase=HTH_7--::--DUF4158=DDE_Tnp_Tn3-> --Resolvase->::--DUF4158=DDE_Tnp_Tn3-> <-DUF4158=DDE_Tnp_Tn3--::--Resolvase-> --DUF4158->::--DDE_Tnp_Tn3-> --Resolvase=HTH_7->::--DUF4158->	

Tn*7*	--Tn7_Tnp_TnsA_N=Tn7_Tnp_TnsA_C-> --Tn7_Tnp_TnsA_N->::--rve=Mu-transpos_C-> --Tn7_Tnp_TnsA_N=Tn7_Tnp_TnsA_C->::--rve=Mu-transpos_C-> --Tn7_Tnp_TnsA_N=Tn7_Tnp_TnsA_C->::--HTH_28=rve=Mu-transpos_C-> --Tn7_Tnp_TnsA_N=Tn7_Tnp_TnsA_C=rve=Mu-transpos_C-> --Tn7_Tnp_TnsA_N=rve=Mu-transpos_C-> --Tn7_Tnp_TnsA_N=Tn7_Tnp_TnsA_C=rve->	

Table I.3. Statistical significance for the non-random orientation of IS elements in the Phyla Bacterioidetes, Cyanobacteria and Spirochaeta

	Bacteroidetes			Cyanobacteria			Spirochaeta		
	Chr.	*IS*	*Orient.*	*Chr.*	*IS*	*Orient.*	*Chr.*	*IS*	*Orient.*
IS1	7	29	0.426	7	49	0.283	-	-	-
IS3	28	171	0.333	-	-	-	13	50	0.894
IS4b	11	48	0.206	-	-	-	-	-	-
IS4c	-	-	-	4	26	0.254	-	-	-
IS5a	16	112	0.775	-	-	-	5	28	0.656
IS5b	10	68	0.630	13	142	0.470	-	-	-
IS5c	-	-	-	2	22	0.312	-	-	-
IS5d	5	62	0.199	11	154	0.556	-	-	-
IS21	14	62	0.393	-	-	-	-	-	-
IS30	6	23	0.812	-	-	-	6	24	0.869
IS66a	9	30	0.513	-	-	-	-	-	-
IS66b	2	21	0.263	2	22	0.663	-	-	-
IS110	38	151	0.119	6	26	0.398	8	116	0.577
IS200	36	132	0.546	14	59	0.825	-	-	-
IS200/IS605	-	-	-	11	45	0.459	-	-	-
IS256	17	68	0.265	-	-	-	7	66	0.0603
IS607	-	-	-	8	39	0.833	-	-	-
IS630	7	71	0.533	16	392	0.884	-	-	-
IS701	3	11	0.551	11	85	0.855	-	-	-
IS982	17	104	0.676	-	-	-	-	-	-
IS1182	20	78	0.926	3	23	0.583	-	-	-
IS1380	-	-	-	4	22	0.833	-	-	-
IS1595	16	88	0.263	-	-	-	6	24	0.124
IS1634	-	-	-	5	46	0.518	-	-	-
ISAs1	9	28	0.147	6	34	0.366	-	-	-
ISAzo13	-	-		5	29	0.359	-	-	-
ISL3	23	81	0.764	8	73	0.775	-	-	-
ISNCYa	-	-	-	19	102	0.247	8	21	0.572

Figure I.4. Sequences of transposase regions containing the β motif. *Escherichia coli* peptides used for biochemical analysis (Fig. 4.6 and 4.7) are boxed, and residues putatively corresponding to the β binding motif are in bold. Phylum code: Actinobacteria (A), Bacteroidetes/Chlorobi (B), Chlamydiae/Verrucomicrobia (V), Cyanobacteria (C), Deinococcus-Thermus (D), Firmicutes (F), Fusobacteria (U), Nitrospirae (N), Planctomycetes (M), Proteobacteria (P), Spirochaetes (S), Tenericutes (T), Thermotogae (H) and Archaea (Arch.).

IS*5a*

P *Escherichia coli* (AAB53644) 305-RESQIQGVAENDN**QL**AM**LF**TLANLFRADQMIRQWERSH*
P *Yersinia enterocolitica* (CAE46781) 231-TKVRFRGLVRNTA**QL**V**TLF**ALSNLWMARRHLLASAGEVRL*
P *Vibrio cholerae* (EKG68054) 289-IKARYKGLMKNDN**QL**AM**LF**TLANLVKVDQLIRRQARSA*
P *Pseudomonas putida* (CAB99195) 286-VKTRFRGLAKNTA**QL**V**TLF**ALSNLWMARRHLLTNAGEVRL*
P *Legionella pneumophila* (CBX00305) 308-IVICVMVVNQALV**Q**SMD**L**TAYRY*
P *Neisseria meningitides* (CBA09446) 333-ARAAYFGLSKVSA**Q**SH**L**KAMCLNLLKAANRLSAPAAA*
P *Ralstonia solanacearum* (AAR23794) 289-RKVRYKGLAKNTA**QL**F**SLF**GLANLVLARRQLLASPGSIAS*
F *Bacillus* sp. (ZP_01722218) 415-RWTTLRGLKKLSM**Q**AM**L**T**F**AAMNLKKLATWTWQVA*
F *Streptococcus thermophilus* (CAE52417) 311-TLTNLLYNIFRFE**QI**KR**L**GLKSWA*
F *Lactobacillus acidophilus* (YP_004292069) 549-LCFYIRAKNRVSS**Q**T**LF**KRKIKLKSTSKLNP*
F *Thermoanaerobacter ethanolicus* (ZP_08211387) 478-KQGLKFYFVNKLV**Q**IK**LF**AFLYKYYR*
C *Crocosphaera watsonii* (EAM49691) 446-ITFLVVNLSKLLR**QL**L**SLF**LSLFTNNRTGNSSNRLSLILIIL*

IS*30*

P *Escherichia coli* (NP_415922) 329-TNGLIRQYFPKKTCLAQYT**Q**HE**L**DLVAAQLNNRPRKTLKFKT-370
P *Yersinia kristensenii* (ZP_04623201) 329-TNSLIRQYLPKKTCLAQHS**Q**EV**L**NQIADELNDRPRKTLKFKT-370
P *Pseudoalteromonas citrea* (ZP_10273944) 331-TNRLLRQYFPKKTSLHGFD**Q**SY**L**DKIANKLNNRPRRILNYLT-372
P *Burkholderia xenovorans* (YP_555985) 331-TNRLLRQYLPHGTQLDHYS**Q**AD**L**NKIAARLNERPRKTLGFRS-372
P *Pseudomonas fluorescens* (YP_348496) 332-TNGLLRQYFLKGTDLAEHS**Q**AT**L**NEVARQLNSRPRKTLDYET-373
P *Ralstonia eutropha* (YP_728612) 316-TNGLLRQYLPKGTDLSVYS**Q**AK**L**NAIARRLNERPRKTLNFDT-357
P *Rickettsia massiliae* (YP_001499437) 274-MNSMIHRILPKNTDITTVT**Q**RG**L**DNVAEILNNMPRKIFGYKT-315
F *Streptococcus mitis* (YP_003445190) 312-NHTVLRAILPKGTSFDQLT**Q**KD**V**NLVISHVNSLKREEFQGKS-353
F *Streptococcus salivarius* (CAA78948) 289-FNGLLREFIPKGCSLKELN**Q**N**LL**EDYTKAINERPRRIHGYQS-330
F *Streptococcus pneumonia* (CCG13893) 313-NHTLIRDILPKGTSFDNLT**Q**ED**I**NLVCSHVNSVKRAALNGKS-354
F *Staphylococcus aureus* (YP_039551) 261-TNGLLREFFPKKTDLAKVN**Q**EQ**L**NYALDSINYRPRKCLNWKF-302
S *Leptospira biflexa* (YP_001840433) 286-FFPKGTDFSKLKKSQIKKV**Q**T**LL**NQRPRKTLNWNTPEEEIRA-318
S *Treponema succinifaciens* (YP_004365962) 319-NGLPRKRLGYKTPEELFNE**QL**D**LI**YRL*
A *Frankia* sp. (YP_001508751) 434-NDRPRKTLDWKTPTEAMNN**QL**L**SL**QQPGVARTG*

IS*66*a (TnpB)

P *Escherichia coli* (YP_424826) 66-LFTKRLERGRFVWPVTRDG-KVHLTPA**QLS**M**LL**EGINWKHPKRTERAGIRI*
P *E. coli* (YP_003235003) 64-LFTKRLEEGQFIWPAVRDG-KVSITRS**QL**AM**LL**DKLDWRQPKTSSRNSLTML*
P *Shigella sonnei* (YP_313273) 66-LFTRRLERGRFVWPVTRDG-KVHLTPA**QLS**M**LL**EGIDWKHPKRTERAGIRI*
P *S. boydii* (YP_407150) 66-LFTRRLEEGQFIWPAVRDG-KVSITRS**QL**AM**LL**DKLDWRQPKTSRLNALTML*
P *Polymorphum gilvum* (YP_004305836) 66-LVHKRLECGKFVWPQAQDG-VMRISSA**Q**MAA**LF**EGLDWRLVRPERARRPLVAG*
P *Ochrobactrum intermedium* (ZP_04679718) 72-LFTKKLERGRFIWPSAADG-TVVITPA**QL**G**YLL**EGIDWRMPQKTWRPTSAG*
P *Methylocystis* sp. (ZP_08074780) 66-LFAKRLEDGEFRWPKIEDG-TMRLSAT**Q**F**S**A**LL**EGLDWKRVHTKETPAPALPG*
P *Agrobacterium vitis* (YP_002540116) 72-LFYKVLERGYFPWPRAKEG-VAPLTQA**QLS**M**LV**EGIDWRRPAWTSAPARTG*
P *Rhizobium leguminosarum* (YP_002973065) 66-LAYKRLEEHTFTWPGIKDG-LMTLTHA**QF**EA**LF**AFDLVVSGHA*
P *Pelagibaca bermudensis* (ZP_01442077) 66-MAYKRLEESTFTWPAIRDG-AMTLNRA**QF**EA**LF**AGLDWRRVRSLEVRRPAVAE*
P *Desulfovibrio magneticus* (YP_002952289) 66-LWHKRLERHVFRWPTR-EAEVLAIDSR**QL**AW**LL**DGLDPLAVTGHSRLEYSTLF*
P *Magnetococcus marinus* (YP_864253) 67-LWQKRLEKDRFHWLRQGGAAEIQITGR**QL**NW**LL**DGYNLAAMKGHNKLHFSSIV*
P *A. ferrooxidans* (YP_002218731) 65-LVYRRLDQGRLHWPRADAG-ALELSAA**Q**WAM**LV**EGRPWTPLPTLEKCTPKLL*
P *Burkholderia glumae* (YP_002907560) 66-LMLKRLEADHFVWPHR-EQAVIELTTE**QL**HW**LL**DGIDIDAMQRHPARRYRHAS*
P *Cupriavidus taiwanensis* (YP_001795972) 66-LFLKRLEADRFAWPR--GAAVATLSVE**QL**HW**LL**DGIDISAVQRHPPRHYQRAV*
F *Streptococcus pneumoniae* (YP_002742797) 67-LLYKRFENGRLTWPST-EKDVKALTPE**QV**DW**LM**KGFSITPKINPSESRDFY*
F *Bacillus pseudofirmus* (YP_003426249) 66-LYYRRLEKGTFPWPEDSSSSPQMISHR**QF**RW**LL**DGLSIDQKSAHPKVTAQRVI*
F *B. selenitireducens* (YP_003698933) 66-LYYRRLEKGRFPWPTSGSDEPMIITER**QL**RW**LL**DGLPLDQKGAHRKMNPEKVV*

IS*66*a (TnpC)

```
P Escherichia coli (YP_003235004)              MSQKYLIRIAELERLLSEQAEALRQKDQQLSLVEETEAFLRSALTRAEEKIEEDEREIEHLRAQIEKLRRMLFGTRSEKLRREVELAEALL-----KQREQDSDRY
P Shigella dysenteriae (EFW49566)              MNQKYLIRIAELECQL-------RQKDQQLSLVEETEAFLRSALARAEEKIEEDEREIEHLRAQIEKLRRMLFGTRSEKLRREVEQAEALL-----KQREQDSDRY
P Yersinia pestis (YP_003566509)             MVMSQDYLARIAALEDAL-------RQKDNQLSLVAETESFLRSALARAEEKIENEEREIEHLRAQIEKLRRMLFGTRSEKLRRQVEEAEALL-----KQQEQQSDRY
P Klebsiella pneumoniae (YP_001687996)         MNHDYLARIAALEDAL-------RQKDSQLSLVAETESFLRSALARAEEKIENEEREIEHLRAQIEKLRRMLFGTRSEKLRRQVEEAEALL-----KQQEQQSDRY
P Pantoea vagans (YP_003933479)                 MKRSLSAENDRLRALL--------DTQQRSLEQMAEYNRLLS-----RRVAAYASEINRLKALVAKLQRMQFGKSSEKLREKTARQVREAEE---RISGLQEEMA

P Escherichia coli (ZP_07592975)               MDISLLSTTR-------DPEQLRALAIAMVQKAMAESQNLANVVQEKDRNIAELQNRIRILEEQMKLARQQRFGKKCESLAG---MQRSLFEE-DVDADIAEISAH
P Shewanella benthica (ZP_02157382)                                  MASQYSEIAELKQS---------------VQRLLEPFRLAQQQRFGASSESHN----YQGELFNE---AEVTLD-EPE
P Fulvimarina pelagi (ZP_01440753)             MATALEALPD-------------DPGTLKAMLIA--ERVRSE-----------------RLEQIIKELQRHRFGRRAETLPE---DQLLLGLE-DVEQGVAVEEAE
P Magnetococcus marinus (YP_866522)            MKIKPQTLPD-------------DPAELKALVQSLQEEMKL-------------------LREQLHILISKRFGRSSEKYDP---NQLGLFDEAELIGATAADDVE
P Aliivibrio salmonicida (YP_002262549)           MIDKIKP--------------LPDTIDELKALVLQLENK---------------YNRLLEQFRLAQHQRFGKSSESDS----TQFDLFNE---TEEEIIIEND
P Brucella ceti (ZP_03787083)                  MLNRGQHLPR-------------DPDILVGMILERDAEIER-----------------LKVLLKAANAKPFGQRSEQLAHMVERQIRLDLG-DVVHEPEVASAE
P B. cellulosilyticus (YP_004096872)           MRPMEKIKNDKDAQF--------YKERAEKLELEKEELEAK---------------LKWYEEQFRLSQKRQFGASSEQTT----GQLSLFNE---IEDTSNKDVE
P Methylocystis sp. (ZP_08074907)              MSRAAADLPE-------------DPAELRRFAEALAAEVHAK------------TLLIEKLKMQLAVLRRARFGRSSEKLDRDIEQLELLIG-DMEESDAERQAP
Polymorphum gilvum (YP_004305837)              MDAAVLAREN---------------ALLKARLIEVEAALAES------------QEANRRLEDILRTSQREKFGKRSEKLSP--DQFNLPLE-DAELAQGVLEAM
Planctomyces maris (ZP_01851762)                                    MIHQLG---ETVGE-----------QQREVEQLKHFIDRLLRQRFGARSEKIA---PNQMSLFDE--PEAAEEATDPE
B Parabacteroides merdae (ZP_02032857)         MIHTDTMELIIKNQQEQIKGLLETNRTLVESNQKLMEQTGE-----------LQQKVQELLSQVAWLNRQLFGRKSEKLASLDPNQLALFDT--LANPRQEE---
B Bacteroides salanitronis (YP_004258102)      MIRQDTMEQIIRSQQEQIAGLLETNRSLVESNGKLLEQTDA-----------LQRKIQELLSQIAWLNRQLFGRRSEKLAALDPNQLSLFDS--VPATGQDEDIR
V Chthoniobacter flavus (ZP_03131569)           MPDSTFPN-----------------AAELLARIAELEKE---------------NALLRQKIDALARKIYGVSSEKLDP---AQLHLLLQ---GLDEPGKAPE
F Lactobacillus casei (YP_005858453)                                 MSAEAVVTTEQFEYLKQE---------------NALLREQVEFLMRRLYGTKRESLTD--GQVDLFDQ--TKTFVAPTVPF

F Streptococcus pneumoniae (NP_358890)                                        MKIIQQQSAIIDSLTNELSLLCEQVAYLTQKLSGKSSEKSVCPFGQLNLFEEESPSEKDGDVPS*
F Lactobacillus parafarraginis (ZP_09394150)                        MILSLKVKQRTVTQLEIDELRKENAELRALVAKQAKQIELLQEQVNYLMSKLYGKSSEQTPEDGQTSLFEDDENGVFEQPESTGE*
F Enterococcus faecium (EJX82788)                                      METTDTLLQLLQEAHKTNQAQQQTIQNLTTEIQLLNEKVNYLTNKLFGRSKETLFEETNGQLNLFSDEEISVSVPEA-77
```

IS*91*

```
P Escherichia coli (ACO24927)              463-IEDPKVIEQILKHLKQKTAKANAAKQRELPPERAPPLTPSLFDPSQSRLFD*
P Gamma proteobacterium (ZP_01617481)      453-IEDPSVIKKILEHLDAKSMAL--TSANQLPEPRAPP-QAELFD*
P Gamma proteobacterium (CBL46704)         464-IEDPAVIEKILQHLAMKESLP----LPRVHEARAPPDQAALFQL*
P Nitrococcus mobilis (ZP_01126307)        435-IEDPEIIEKILAHLDAHVTEP---EATRRPPCRAPP-QRGLFDETG*
P Aliivibrio salmonicida (CAQ77725)        357-GYVKVDPYECILCESRLVFTNFRVGNSVNDLVTHAIVQSELRAA*
F Lactobacillus casei (YP_806573)          394-RHYMLEVNQNIAKEAYQTKYKQAEAAAYDRCRFSWERQRRIYLSEMPQA*
F Clostridium hiranonis (ZP_03294328)      154-IWHYKYGLIYNVLDKSNYKRIIYEEIIEKEISLNTTTQKELF*
M Kuenenia stuttgartiensis (Q1Q731)        249-IYNEIEEIMRGKYEPKEEKVIKPEGDGGTIRPTPRRVQIPLFSM*
```

IS*200*

P *Escherichia coli* (ZP_03029803) 125-WSWGYFVDTVGV--NEEIIRRYARYQEKMEQTHEQ**Q**ME**LL**E*
P *Vibrio cholera* (AAC01554) 106-WARGYFVDTVGV--NEEIIRRYVRHQDKKELEPEQ**QL**E**LL**RD*
P *Shewanella baltica* (YP_001364938) 106-WARGYFVDTVGV--NEEIIRRYVRHQDKQDQEHEA**QLSL**QMM*
P *Shewanella woodyi* (YP_001761123) 107-WQRGYFVDTVGI--NEEVIRRYVKHQEKVEKQEQP**QL**D**L**K*
P *Serratia odorifera* (ZP_06639291) 138-WARGYCVDTVGI--NEEMIRKYVKYQEKHEVE-ES**QL**P**L**KEV*
P *Nitrosomonas* sp. (YP_004295562) 107-WTDGYYVATVGERADWGEVERYVKNQGKPKEE-LR**QL**E**FF***
P *Escherichia coli* (NP_752024) 123-WCRGYYVDTVGK--NTAKIQDYIKHQLEEDKM-GE**QLSI**PYPGSPFTGRK*
P *Yersinia pestis* (YP_005509676) 123-WCRGYYVDTVGK--NTARIQEYIKHQLEEDKM-GE**QLSI**PYPVSPFTGRK*
F *Streptococcus pneumoniae*(ZP_01829269) 107-WCRGYYVDTVGR--NQKVIAEYIQNQLQEDRV-AD**QLTLF**ESVDPFT-150
F *Clostridium botulinum* (YP_002650746) 105-WCRGYYVDTVGR--NKKAITEYIKNQQKEDMI-SD**QISL**KEYMDPFK-148
F *Bacillus halodurans* (BAD18222) 106-WCRGFYVDTVGR--NKKQIQEYIRNQLREDYM-GD**QLTLF**EEYDPFT-149
F *Enterococcus faecium* (ZP_05832064) 105-WCRGYYVDTVGK--NAKKIEEYIANQLQEDLE-YD**Q**M**TL**KEYIDPFT-148
F *Ruminococcus albus* (YP_004103952) 108-WCRGYYVDTVGK--NKKKIAEYIRNQLQEDIV-CD**QISLF**ETVDPFT-151
F *Staphylococcus epidermidis*(YP_188749) 107-WCKGFYVDTVGR--NKKVIENYIRNQLQEDIV-AD**QIS**MEEYLDPFT-150
F *Filifactor alocis* (YP_005053889) 110-WCKGYYVDTVGK--NTKAIQEYISNQLKVDRE-SD**QLSIF**DPRDPFTGSK*
U *Fusobacterium nucleatum* (ZP_00143858) 105-WCRGYYVDTVGR--NKERIAQYIKNQIEEDKI-MD**Q**M**TL**KEYFDPFN-148
P *Xenorhabdus nematophila*(YP_003713015) 106-WTDGYFASTVGKHGDEQMIGRYVQNQGKKYHK-LHSDH**QL**A**LF***
P *Psychromonas ingrahamii* (YP_944951) 107-WTQSVFVETIGNATEE-VIRKYVQNQLIELDRKEINSD**QL**D**LF***
C *Microcystis aeruginosa* (CAO87910) 106-WSDGYYASTVGKHGDEGMIARYVKEQDKEYLQ-LHQNL**QLSLF***
C *Acaryochloris marina* (YP_001519691) 104-WSRGYFVSTVGR--DEEVIRRYIRHQEQEEQK-LE---**QL**N**LF**RA*
B *Pelodictyon phaeoclathratiforme* (YP_002018173) 106-WTDGYYAGTVGKHGNEDMIGKYVKGQGGTYQK-RYSDY**QLSLF***

(Arch.) *Methanosarcina barkeri* (YP_307176) 106-WSDGGYIGTVGDGTTSDVIKSYIENQGNQEEKEAYK**Q**MK**I**ID**F**Q*
(Arch.) *Methanosarcina mazei* (NP_632811) 106-WSSGKFYRSVGNVTAD-TIKHYIKESQKKPKTEVKSSKSAKPD**Q**RI**I**DD**F***
(Arch.) *Methanosaeta concilii* (YP_004384718) 106-WSSGKFFRSVGNVTADTIQ-HYIKESQGKPKAESKVCRSRESG**Q**RR**L**DD**F***
(Arch.) *Sulfolobus islandicus* (YP_002831273) 108-WSPSYFLATSG**Q**VT**L**EV**L**K-KYVESQGKE*
(Arch.) *Halobacterium salinarum* (YP_001688288) 106-WQPGYFLATTG**Q**VS**I**DT**L**M-DYVDDQ*
(Arch.) *Haloarcula marismortui* (YP_134246) 108-WNDSYCLISTG**Q**VS**L**DV**L**K-QYVEDQRE*

IS*200*/IS*605* (OrfB) Cyanobacteria

Cyanothece sp. (YP_002371738) 367-DLNHGLLAGTAPNFMNTQKERIGEYI**QLSLF**DPTLFGG*
Microcystis aeruginosa (YP_001656728) 371-IENRGKNAVGLTVLENACGGDLTGVV**QL**N**LF**DLVKSLRTKN-411
Arthrospira platensis (BAI93705) 374-NGCGERVRLSVKKAHLNEASTRPAFQ**QLSIF**DLLK*
Synechococcus sp. (ZP_05040300) 360-ASVMGVVASADGVFDNPLNSMNQSAQ**QL**T**LF**PMSA*
Synechocystis sp. (NP_441190) 368-KNGRGGKRQTTSVAASGEASTHRKAI**QL**T**LF**AS*
Trichodesmium erythraeum (YP_721194) 404-ALSKGKTCRKKPANCDEMLTRFESFK**QL**N**LF**D*
Microcoleus chthonoplastes (ZP_05026728) 410-ALVGLSINQPGGTGLSCKLSRTIKYV**QLSLF**DDFRATKNPDLSS*
Crocosphaera watsonii (ZP_00513803) 435-TPKPELTGSSHRETSVSLEIEPGNPQ**QLSLF**EWVNGEVIPC*
Acaryochloris marina (YP_001521276) 266-NGCRRECKSEVNSAVLSDASTRLVDK**QL**E**LF**AS*

IS*1380*

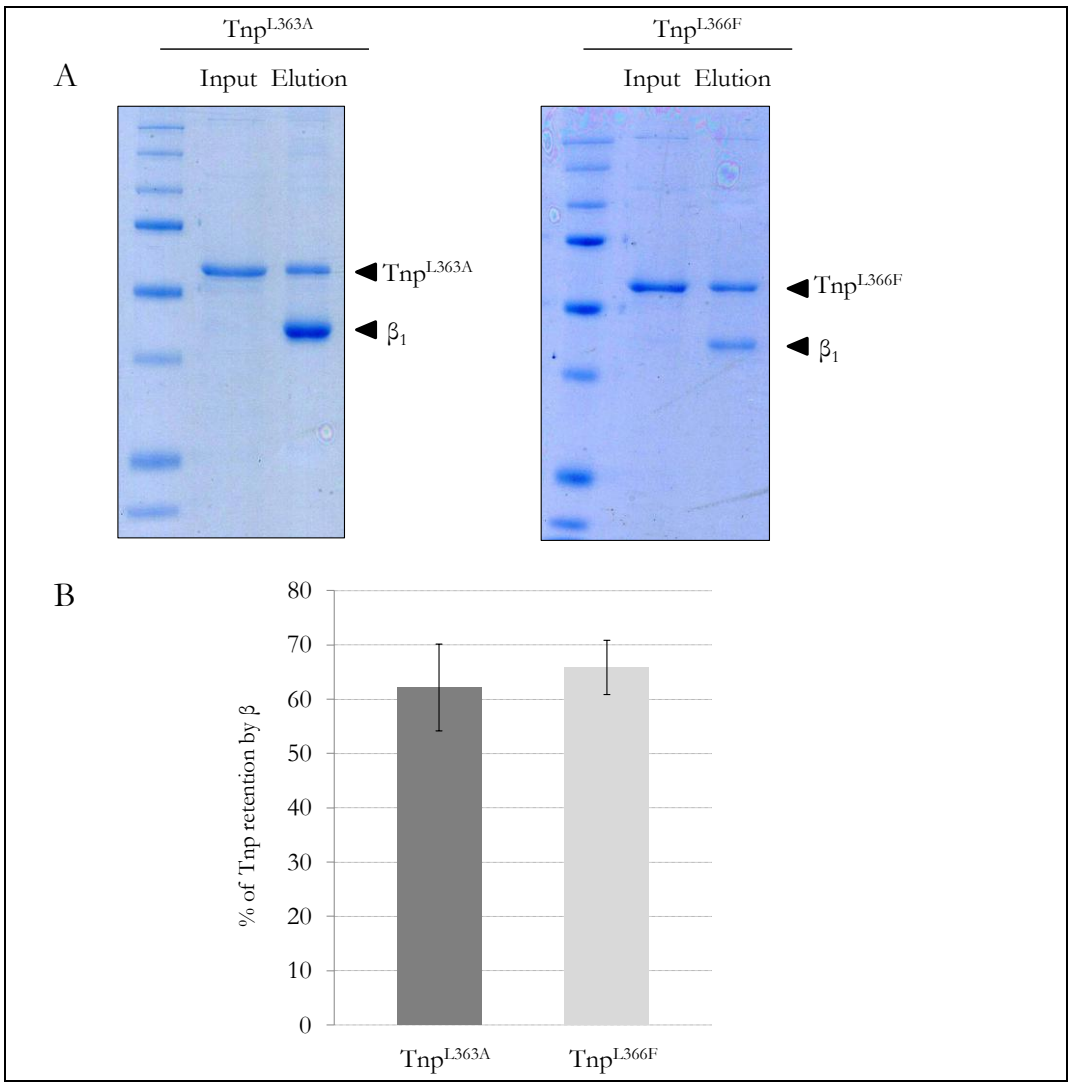
P *Escherichia coli* (YP_003829282) 271-WEKDRRFVVSRVLKPEKE---RA**QLSLL**EGS----EYDYFFFVTN-TTLLSEKV-316
P Gamma proteobacterium (ZP_05061507) 293-WSRSRRFIAVRRLAKVKK-EGPQ**Q**S**LII**EPV-----YDYFCYVTT-ERLTPWQA-339
P *Azoarcus* sp (YP_195484) 286-WKKAKRTLRLVVRVTERTIDKKG**Q**H**LL**APEIEIEGWWTSLD-------VPMADV-332
F *Bacillus coagulans* (YP_004859304) 280-WEKPRRVAVIRKADKYEE----D**QL**Q**LF**DF-----LWDYEAIVTT-MDWEPMDI-323
F *Enterococcus casseliflavus* (AAX38177) 270-WEKDRRFVVSRVLKPEKE---RA**QISLL**EG----SEYEYFFFVTN-TTLLSEKV-315
F *Desulfotomaculum* sp. (YP_001111807) 316-WSKARRFVFIRETQEPKV--SGE**QL**N**F**-DL----KTFDYQVIITSSDEYNPEEV-362
F *Geobacillus* sp. (YP_002949920) 289-IDGNTYTYVQVTQVTERTMERNG**QL**M**LV**PDYEVES-YWVRLKGY--EHVRMSDV-339
B *Tannerella forsythia* (YP_005013031) 278-WQKPRRIVIVRQKIEKRPQAGGK**QLSLF**PEDEIHRNYRYSAYFTN-QTCSMVDV-330
B *Prevotella buccae* (ZP_07884135) 291-PRRIVMVRQEVEKRPKAAGKQVR**QL**E**LF**EDEQDFGKYRYSCFVTN-LALPAKIV-343
A *Rhodococcus jostii* (YP_707285) 314-HQIPGR-LVVRRIPDLRPPKDQG**Q**GT**LF**DI----WRFHAFFTTTDPDDLDTVDA-362
A *Mycobacterium gilvum* (YP_001133130) 278-RGWPAGMRVIARR--ERPHPGA-**QL**R**L**TDDNG--WRITCFATNTP--GWSIADL-323

8.2. Appendix II

Figure II.1. Tn*5* transposase amino acid sequence.
Amino acids truncated in $Tnp^{\Delta C20}$ are in bold type, underlined in Inh and boxed in $Tnp^{\Delta N7}$. Amino acids L363 and L366 that are mutated to alanine in hyperactive mutants (Tnp^{L363A} and Tnp^{L366F}), are marked with an asterisk.

```
         10         20         30         40         50         60
MITSALHRAA DWAKSVFSSA ALGDPRRTAR LVNVAAQLAK YSGKSITISS EGSEAMQEGA

         70         80         90        100        110        120
YRFIRNPNVS AEAIRKAGAM QTVKLAQEFP ELLAIEDTTS LSYRHQVAEE LGKLGSIQDK

        130        140        150        160        170        180
SRGWWVHSVL LLEATTFRTV GLLHQEWWMR PDDPADADEK ESGKWLAAAA TSRLRMGSMM

        190        200        210        220        230        240
SNVIAVCDRE ADIHAYLQDK LAHNERFVVR SKHPRKDVES GLYLYDHLKN QPELGGYQIS

        250        260        270        280        290        300
IPQKGVVDKR GKRKNRPARK ASLSLRSGRI TLKQGNITLN AVLAEEINPP KGETPLKWLL

        310        320        330        340        350        360
LTSEPVESLA QALRVIDIYT HRWRIEEFHK AWKTGAGAER QRMEEPDNLE RMVSILSFVA

        370        380        390        400        410        420
VRLLQLRESF TLPQALRAQG LLKEAEHVES QSAETVLTPD ECQLLGYLDK GKRKRKEKAG
  *  *
        430        440        450        460        470
SLQWAYMAIA RLGGFMDSKR TGIASWGALW EGWEALQSKL DGFLAAKDLM AQGIKI
```

Figure II.2. Tnp^{L363A} and Tnp^{L366F} interact with β sliding clamp.
A) Pull-down assay where β is covalently coupled to magnetic beads and proved for interaction with Tnp mutants. Coomassie stained SDS-PAGE gels of the interaction experiment. In each gel, first lane represent total amount of Tnp used in the assay (input) and second lane the eluted products (elution). From left to right, Tnp^{L363A} and Tnp^{L366F} are able to interact with β B) Densitometry of each Tnp input and eluted band are plotted as percentage of Tnp retained by β. Presented data are the result of three independent experiments, and standard deviations are also shown.



8.3 Appendix III

Table III.1 Microarray probes and log_2 (ratio) values for the set of 85 IS-associated genes identified in the *Acidiphilium* sp PM genome.

Each IS-related ORF (transposases, accessory proteins) was classified into IS families (See Materials and Methods). With the exception of five cases, each ORF was represented by two oligonucleotides, and spotted in triplicate in the array. SD, standard deviation.

	IS Family and ORF number	Oligo #	Oligonucleotide Sequence	Log_2 ratio (2011/2007)	SD
1	IS4-4	473 474	GAGATCCTGGCATTGGAGCGCAGCACGAAGACGCC TGCAGGATCAGCGCGAGAAAAGAACAGAACACATGGC	-0,022	0,301
2	IS4-5	383 384	GCTTCCTCGGGGTTCAGGCAAAGGACATAACGCTTG GGCGTTGATATTCCGAGGTTTTTCAGGATCTCGGC	-0,530	0,444
3	IS66-9	1237 1238	CAACCCCAACTCCATCATAAACCCGAACACAGGTCCCGTC CCGCACTCGCCATTACCCGCACCGGAATGAATTCTGGTAC	-0,045	0,120
4	IS110-1	1273 1274	CATAGATCTTGCCGTGCCGCAACAGAAATCCGGTGAGGTG CTTCTTCGGAATCAACGATGGCGCAACCACCGTGCAGTCG	0,295	0,064
5	IS110-2	1343 1344	GAATGCTCCGACGGCACAAGCCCCACATAGGCCATCAATT GCATCCTCAACCGCATTGACATAGTCCTGCAGAACGATCT	0,441	0,012
6	IS110-3	1231 1232	CAGGGTCCAGTGCTTCCACGCGCTGAAAACAGTGACAATC GAGATCGAGGACCTGCGGGTAGTATCGCCAAAGTTGCTGA	0,026	0,056
7	IS110-4	853 854	CGTTCTTGTTGCGTTTGACGAAGGGCTTCACATACTGCGG CTCAGCCTCATCGACCCCGTGGATCGTGAAGACATGCTTG	-0,457	0,099
8	IS110-5	847 848	ACTCTCGAAATTTGCCGGGTTTACCGACAGGGCTATGGTA GATCTTCTGATCGAGTTCAGCGACATGCCGTCCCATCTGG	0,440	0,007
9	IS110-6	849 850	GACATTTCCTTTGCCTTTGGCCGCGACGATACCGAACTCC TTTGACGAAAGGCTTCACGTATTGCGGCGGGATCAGCTTC	0,112	0,045
10	IS110-7	181 182	CTGATTGCTTCGGCATCATTTGCGTCGTTCTTCTGGTTGC GATAATCAGCAAACGCCGCAAGGTTCGCTCGCCCATCTTC	0,300	0,021
11	IS110-8	1235 1236	CTTCACGAAGGGCTTCACATATTGCGGTGGGATCAACCTG TTCCTCGGCAGACAGGAGAACAAGCAATGCTGAGACGTTC	0,260	0,014
12	IS110-9	161 162	CAACTGTCCGGTCTCGAGCCCGACCCGAAC CCCTCTTGCCCCGCCAGACCAGCTTCCCGT	0,302	0,184
13	IS110-10	1323 1324	CAAACAGCCAAGCGGTCGCTGATTTGCTGCCTGGTTTAGC TCATTGCCTGCCTTACTTATCTTTCCCATACGATGCTTGCCG	0,250	0,134
14	IS110-11	23 24	CAAATAACGATTGCCGCGTTTGCTGATGCCGAGCAG CGCCAAAACGGCCCATGCGATCCGAGCCAGCTTCG	0,040	0,071
15	IS110-12	1327 1328	GTCGTTCCTTCTCGTCCACCGCGTGGATCGTGAAAACATG GTGAGTGACCAGCTCGAAATCTGGATGGATCAGCTTCACC	-0,415	0,092
16	IS1380	145 146	CGAGCGTTTCAGGGTTTCGAGGCGGACGGCGAGTTTGATC CGCCATCATTGAACTCTTTGAACCGCCGGATTTTCTCTCC	0,213	0,222
17	IS1595	901 902	GAAATCCTTGAGGCGGGAATGGCGGTTATTGACGGTCTGG GTTGGGATGAAGGCCGATGAGGTGAAACCAGCGAAGATAG	0,136	0,120
18	IS1634	535 536	AGCACCGGCCTTCAAGATAGGACGAGCTGACATCATAGAG CCACTCGACGTAATAGGCAATCATGCAGAGAAAGACATGCG	1,612	0,046
19	IS21-1	239 240	ACGGGATCGCGGTGAAGTCGAAGGTATCGAGGCTT CGGTGCCGGAATTTCCCAGAGCGATGATGTTGTCG	-2,063	1,582

20	IS21-2	1436	GATCGAGGAGTGCGGTGGTCATCTTGGCATCGCCGAAGAC	0,299	0,100
21	IS21-3	1021	CAAGAGCATTGACCAGATCGACGGTGGCAAATAACCTGAT	-0,050	0,130
22	IS256-1	1295 1296	CTGTTCTCCGCCTCCGCTTCGTCACTCAGATGGTAATCCA TCTTCGCAATAAAAGGTAGAGGCTTGACCCACATTTCTCACGC	0,334	0,066
23	IS256-2	363364	CATCCCCCGCGCATACATCGAGATGATCTTGTCATCGAAG CTGTTCTCCGCCTCCGCTTCGTCACTCAGATGGTAATCCA	0,352	0,201
24	IS256-3	951 952	CTCGCGCACCTTCACATAGGTGGCATCGAGCCAGAGATAG GTTGGGGAAGATACAAACGACATCGGCACGGCGCTTGATC	0,016	0,008
25	IS256-4	1475 1476	CCGGTCAGCGCCTTGACTTCCGTCTCCATGATCCGTTCGG CTTCTCAACCAGCTCGATCAGCGCCATTCTCTCGTCGGTC	-0,133	0,137
26	IS256-5	977 978	CATGGATTTCGGAAAGTGGTTGAGCACGTTGGCGATCTTG GTAGTCATAGAAGGCCAGCAGAACCTCGGTGTCCTTGGTC	0,385	0,098
27	IS256-6	113 114	CTTGAGATAGGTCGCATCCAACCACAGATACGGCCACTCG GATCAGCCCATCCACATCCGTCTCCATCAGCAGTTGAACC	0,395	0,219
28	IS3-1	1107 1108	CAAAACGACCTGAACGACTTCCGCTGACCAAAATAAGCCC CACTATGCCACTCACACACGAAATTCAGGACACTGCCGTC	0,295	0,375
29	IS3-2	415 416	CTTCCACAAAATCGTAGGACCACACATGGCCGGGATACTC TGATGTAGAGCGTCTTCACACCGACCTTCCCGAGCCATTG	0,132	0,118
30	IS3-3	1203 1204	CGGGCGATGAATTCGGGACCATTATCCGACCGTATATGCG GTTGAGCAGCTTATCACGCAGCGAGCCATTGAAGGATTCG	0,109	0,078
31	IS3-4	241 242	CCCCGAACTTTCATCCGTGTGCGATTGGAGTCTCG GGCATGATCGAGGCCGAAGAACAGGCTCTCGTTCA	0,289	0,155
32	IS3-5	1269 1270	CTTCCGCTTCCGCCACAATCCGTAGCTTGTCCTCATCACG TATCTTCAGCCCCTGCAATCATCACGGCAGCAGACGGCTC	0,540	0,026
33	IS3-6	1379 1380	CCGTTGATGGTTTCAGAGAGGGCATTGTCATAGGAATCGC CGAGGAAGAGGCGGTTGTTGATCCAGTCCACCCATTCGAG	0,346	0,078
34	IS3-7	959 960	ACTTCATCTGGTCGATAGCGGCGAAATATCGTTGTTCGGC ATCAGGCGTTCGGTGTATTTGATCGAGACATACTGGCTGC	0,132	0,086
35	IS3-8	1497 1498	CAAGCGGCGATCTTCATCTGGTCGAGTGTGGCGAAATATC GAGAAGGCGTCTGTGGTTGAGCCAGTCGACCCATTCGAGG	0,307	0,303
36	IS3-9	265 266	GATACGCCGGGCGAATGTATCGATGACGAACGCCA TCGTGAAGAGGCTGTTCCAGGGCATCGAGGACGAA	0,157	0,174
37	IS3-10	1299 1300	CTTCAGCTTCGCAGATACTTCCGACGGAACACCAGGCGCA GGACTTCGGCAGAAAATTTGTTCGTCGTCGGACTCGTCAT	0,198	0,045
38	IS3-11	1103 1104	CGGCAGGTGGTTCAGCAGGTCTTCATAGGCATTGATAAAGG GTACCAGGTTCGGATCGTGTCATCATCCAGCAGCAGCACC	-0,480	0,033
39	IS3-12	1411 1412	AGACCTTGCGGACACCATAGACCTGGAAATTCTCTTCCCA GTTTGGGTCAGTACGACGTGCGGCATGCTCATAGTAGGTC	0,019	0,064
40	IS3-13	1417 1418	CAACTTCAGGCCGCCATACTCCTTGCGCCACCGATAATAG GAACAATCTCAGCCTCGCGCAACTTGCCGATGATCTCTTC	0,207	0,217
41	IS4-1	73 74	GCAGAAGAAATTCTCGATCAGATGACGCCATTTGTAGGTTTCGTGG TTGCTGTCGAATGCCTTGTCTGCGATCAATCCACCGAATT	0,358	0,007
42	IS4-2	201 202	CGTCAGCCTGGTGAAAGATCGGCCTTGTATCGAAGCTGGC CTGGAGCCTGTCGAGATCGTTCAGCAACTTCTGCCATTCG	0,289	0,074

43	IS4-3	1133 1134	GAAATACAGCGAGGTGGTGTCCATGAACACGACCGAGAGG GTGCCGGGAATGTCGTAATCGTCGAGCCATTTGTCGCAGG	0,200	0,048
44	ISAs1-1	89 90	CAGGATTTCGGGTTGATCGTCGCTGCTGGACAGGAAGTAG GAAACGCCCGGTATCGATCAGCATGAACACACGTCGGAAG	-0,150	0,014
45	IS5b-7	1347 1348	CTCACCACTCTCGGTCGTCTTGACGCTCTGACTGTCGATC GAACTGTCTGTCGGCCTTTCGCACCACCTCGATGCGGATT	0,245	0,077
46	IS5b-8	13491350	CAGGAAGTTGGCGGCCTTCTTATCGTAGCGAGTTGCGATGC CAGATAGTCTCTGATGTCGTCTGCGTCATAGGCTTTGTCG	0,704	0,005
47	IS5b-9	1407 1408	GAGGAAGTTTCGTGCGGTCTTGTCGTAACGTGTAGCGATG TCATCGGCGTCGTAGGCCTTGTCGGCGATCAGGTCTTCTG	0,400	0,010
48	IS5a-1	973	GGGGTGGATGGTATTGGGGAGATTCGGAGGGCTTATTTTG	0,304	0,130
49	IS5a-2	227 228	ATCGCTTCGTCGGACAGCCCATCCATGTGCTTGAGAAGAT GGAACAGTTCGAGATCATCCTGCGAGGCATTCAGCTTCGG	-1,026	0,127
50	IS5a-b	1233 1234	CTCGATGAGATTGCGCTTTCTGTAGGCGACGGGGTCGAAG CTGATCCATGAGGTGACGGCGAGGACGAGACAGATACCTG	0,133	0,081
51	IS5b-1	1001 1002	CCGCAACCAAGGATACAGGGATTTGACGGATTTGAGCACG TCGACAGCATGTAGAAGATGGCCTCCACAATCCTTCGCAG	0,190	0,127
52	IS5b-2	1515 1516	CTGCCCATTGCGTGTCGTTCAGCCAGAATAGCTCC GAGCCCTCGCGAAAGCGATGCAGGATGCCGCTCAG	0,295	0,115
53	IS5b-3	1117 1118	CGTCGAGATCGGTCTTGAGGTGAATTTTGGTCGGGAAGCC TACAACCGCTTCGGGAAGAATTGGCCTCTTTCCTTGGAATTT	0,377	0,222
54	IS5b-4	1479 1480	AATCCCGCCTCGCAGGACGTAGAAGATCGCGTTGACAATC CAGATGATGGTTGATCCGCTCCCAGGTGCCATTGTCACGG	0,014	0,020
55	IS5b-5	1131 1132	GAACACGTCAAGGCGGCTCCAGCGGATAAACTGATTGTAG GTGAGGCAGCGGCTCTTTCCCTCTCCTGATTGATGGAATG	0,131	0,086
56	IS5b-6	441 442	CGCTTGACGATCTGTATGGTCCATTTGCCGATTTTGGTGAG GTGGTGAACGGTGCGAATTCCTTCGGGATCTGCCACCTCC	0,315	0,007
57	IS6-1	1205 1206	CCGATGGTCCTGCTCAACGAGGTTATTCAGATACTTCGACGAT CGGTAGAGATAGACCCATTCACCCTGGACATTCACGTAGGT	0,019	0,225
58	IS6-2	1389 1390	ACGCTTGGGGCTGAGACGGAAGTCAACGGCATTTC CTTCACGTAGGTCTCGTCCACACGCCACGACGCTC	0,118	0,116
59	IS6-3	1535 1536	CATCGACAGGCCGCGCTCCTCCATCATCTCCAC GTTCCACCGGCGCTCGAATTCCGGCGAATAGTG	0,200	0,005
60	IS630-1	1337	AAATTGTCCATGATCACGAGATCGCCCGGCGTCAGGGTTG	0,270	0,100
61	IS630-2	1239 1240	AATCGGGTTGAAATCCGGCGAGTATGGCGGGATGAACAAG GTTTTCCGACCGTGTCCCAGAGTGCGTTGATCGTTCTTTC	0,385	0,064
62	IS630-3	1285 1286	CAACCACCCGTTCACGCAAATCCAGCGAAAGTACCTTCAC GAAGAACCGCCAGATCGTCGAAATGCCAAACCACTCCGAA	-0,001	0,219
63	IS630-4	1145 1146	GATGATGGGTGCTTTGTGGCTGCCGAGATTGTCCATGATC TCAACATGATGGTGTCGATGATCTGTCCGATTGCACTCCAG	0,004	0,023
64	IS630-5	1229 1230	GAAGCTCGGGAATGAGGACGTACTCAACCCAGGTCTCGAA GGTTGAAGTCTGGGCTGTAGGGCGGCAGATATCGTAACTC	0,154	0,014
65	IS630-6	783 784	ACATTCTTCTCAACCATGTCCAGGAAGCGGCGGAAGTCTG TTCTCCAGCGTCGTCGTTATCACCTCGGCAATCTTGTCAT	-0,070	0,297

66	IS630-7	108	GACATATGCCTCGAACCAGTCGCCGTTGATGGGGCCGTCC	0,293	0,130
67	IS630-8	1405 1406	GATCGGGTTAAAGTCGGGGCTGTAGGGTGGAAGATAGAGC GGTGATGGTGTCGATGATCTGTCCGATTGCACTCCAGAGG	-0,006	0,020
68	IS66-1	1503 1504	GAAACCCCACCTTCGCGTCGTCGTTAATTACGTCCA GGCAACAATCCGCGTCTTCTCCTCGTCCGACCATC	0,193	0,113
69	IS66-2	561 562	CACATTGCTTCTCCTCAAGATCGACCACGACCTCGACCTG GAGTTTTCCAGTTCCAAGGCAGCAGGGTGTCGATTTCGTT	0,321	0,149
70	IS66-3	611 612	CGTCTTTCTCAGCGTGCGATAGGAGAGACAGGTCAGGAAG GGCGATATCCGGGTGGTCGAGGAAGGTGAAGAGATGATCA	0,509	0,142
71	IS66-4	1509 1510	CAAGATAGACCCGGACACCCGCACCCGGCGCGATC GTTCCCGCCAAAGACCTGCTGGACCATCATCGCCA	0,847	0,311
72	IS66-5	1339134 0	CCAGATAAACCGCCCTCGATCAAGGCGTTTCGAATACAGCG GATTTCGCCAGTCGATCCCCTCCAGAAGATAACCGATCT	-0,065	0,049
73	IS66a	1089 1090	CTTCAGCCAGATCGCCATCCCGATATTGCCGACGC GACACCGTCGTGCCAGAGCAGCTTGATCTGTCCAG	0,205	0,205
74	IS66a OrfB	263 264	CATGCCAGATGATCTTCACCAGCGAGCCCGACCTG GCTGCGAACTCGTCAACGAAACCACCCCGTCCTTC	0,256	0,185
75	IS66a OrfC	307 308	CGGAGCATGTAGTTGATCGCCTTGGCAAGATCATGGTGGC GTTATTGAGGCGGGCGGTCTGGATCAGGGTGTAGAGGATG	0,262	0,066
76	IS701-1	273 274	GCCCGATCGATAAAGGCATGGCATGAGCCGACA GGCCGTTGGCAATCTTGCCCGCCGAGCCGG	0,310	0,092
77	IS701-2	687 688	CAATGTATCCTCAACCGACGCACCGCCGATCATCATGTCC GGAGTGTCATTTCAATCTCGCCCACGCCGTAGACGCTATC	0,478	0,110
78	IS701-3	679 680	CCTCGTCGCCAGTTGATCCTTCGCCACTTCTCTTCAATCA GACTGACAATTCGCCGTCTTGCCCAACGCAGAAGCATATT	0,546	0,003
79	ISL3-1	1223 1224	GCATACATCGAGATGATCTTGTCATCGAAGCCCGGAAACCG TGTCGTGGGCGTTGATGATTCGGCTTGGCGGCGTAATTTC	0,319	0,326
80	ISL3-2	119 120	GAAAGAAGTTGAGGTGCCGCCAGCTCAGTTCCTTCCAGTC GCATCGAGGTCAATGAACACGGTGACGTAGCTGTGGCCTC	0,278	0,158
81	ISL3-3	1143 1144	CAAGATATCACAGACGAGTTTGCGACTATGGCCGGTGCCG CGGCATTCAGGAAGGCGTGGCTGGCATTCTCCATGAGATG	0,027	0,057
82	Tn3-1	1209 1210	GTGGGTCCAAGACCAAATTCCTTCAGCCGCTTGACCTCAG TCAGCTCGTGAACGATATCCTGAAGGTCCTTCATCGACCG	0,278	0,219
83	Tn3-2	291 292	GATCGGAGAGATGGGTGTAGAATTTCAGACCCGGTTCGGG GCATAATGTTCGCTCAGCACCTCGTAGGGATCGAAATCTTCC	0,043	0,215
84	Tn3-3	253 254	GAAATCGTGCTGTGGCTGACGTTGTAGGAGCGGGC GGCCTTCGCCCGTTCTCTTCCTTCTCCGGTCCGGG	0,000	0,061
85	Tn3-4	923 924	CAAACCTGTAACCCAGCAGATGGCAGAGACCGAAGACATG GATAGATGGTATTCCAGTGGACGATGGCGGCGGTGATGAG	-0,025	0,095

Figure III.1. Structure of IS*1634* in *Acidiphilium* sp. PM.
A) The insertion sequence (1930 bp) contains 12-bp-long inverted repeats (underlined). The initiation and stop codons of the transposase gene are boxed. B) Sequence of the IS*1634* Tnp. The putative DDE motif is marked with asterisks and the β binding motif is boxed.

A

CCGAAGTTTTTATTTGAAAATCATCTAAACGTGAGGCTCCGTGCGGGATACGGCTCGTTATTAATTTCAGAATGTGTC
CATACATGGGGCTTGCGGCGCGACTCGATTTGTGATTCGCTGGCGTC**ATG**TTCATCGACGTGGTTCCGAATGGCCGCT
CGGCGTCGGCGGTGCTGCTGCGGGAGAGTTTCCGTGAGGGGCGGAAGGTCCACAAGCGCACGATCGCCAATCTGAGCC
AGATGCCGGCGGAGCTGGTCGATGGCTTGCGCGCCCTGCTCGCCGGCGGCTCGGTGGTGGGCGGCCCGGATCAGGCGC
TCGAGATCCGGCGATCCCTGCCGCACGGGCACGTGGCGGCGGTGCTGGGGATGATGCGCAAGCTGGAGATTCCGCGCC
TGCTGGGACGCCAGGTCTCGCGCGAGCGGGACTTGGCACTGGCACTGATCGCGAGTCGCGTGATCGCGCCGGGCTCGA
AACTCTCGACGCTGCGCGGCCTGAACCCGGAGACGGCGACCTCGAGCCTCGGGCAGGTGCTTGGGCTCGGCGTGATTG
AGGAGCGCGAGATCTACGCCGCCCTCGACTGGTTGGGCGCGCAGCAGGGGCGGATCGAACGGCAGTTCGCGAAGCGCC
ATCTGCGCGATGGCACGCTGGTGCTCTATGATGTCAGCTCGTCCTATCTTGAAGGCCGGTGCTGCGAACTCGCTCAAC
ACGGTTATAGCCGTGATCATCGGCCGGATCGGCTCCAGATCGTCTATGGTCTGCTGTGCGATCGGGAGGGCCGGCCGA
TCGCGGTCGAGGTGTTCGAAGGCAACACCGCCGATCCCGGCACGATCGCGGCCCAGGTGGAGAAGCTCAAGCGGCGGT
TTCATCTGAACCACGTCGTGCTGGTGGGCGATCGGGGCATGATCACCACGGCGCGGATCCGCAAGGAGATCAAACCCG
CCGGATTGGACTGGATCAGTTGCCTGCGGGCGGGTCAGATCCAGGACCTCGCCGAGGGGCCGCTGCAGATGTCGCTGT
TCGACGAGCGCGATATCGCCGCGATCGCATCGCCTGATTATCCCGGCGAGCGGCTGATCGCCTGCCGCAATGCGGCTC
TGGCCGGGGAACGCCGGCGCAAGCGCGAGGCGCTGCTTACCGCGACCGAGCGGGAATTGACCCGCATCGTGGCGGCGA
CGACGCGCAAGCGCGCGCCGTTGCGCGGCGCGGCCGAGATCGGCCTTGCGGTCGGTGCCGTGATCAACCAGCGCAAGA
TGGCCAAGCATTTCGATCTCACCATCACCGCCGATCGCTTCAGCTTCCGGCGCAACGAAGCCGGTATCGCCCGCGAGG
CGGCGCTCGATGGCATTTACGTCATCCGGACCAGCGTCGCCGCGGAGGCGATGAGCGATGCCGACACCGTGCGGGCCT
ACAAGGACCTCTCCCGGGTGGAACGGGCGTTCCGAACCCTGAAATCGGTCGACCTCGCAATCCGTCCGGTCCATCACT
GGCTCTCGCCGCGGGTGCGCGCGCATGTCTTTCTCTGCATGATTGCCTATTACGTCGAGTGGCATCTGCGCGATGCCC
TCAAGCCGATCCTGTTTCAGGATCACGATCCACTGGCCGCCGAGGCCGAGCGCGCCTCCCCGGTTGCCCCTGCCACGA
TCTCACCCGCCGCCAAGCGCAAGCGGGGGCGACGCCGCAACGACGACAACCTGCCGCTCTCAAGCTTTGCCGACCTGA
TGGCGCATCTGGCCACCCAAACTCTCAACACCGCCGCGCTGCCCAAGGCGCCCAATGCGACCTTCACCACCCTGGCCA
CGCCAACCACACTACAGGCGGCCGCCTTCAACCTCCTCGAAATCGAACCCATGCGTGTCCAG**TAG**ACGTCAAAAGCGG
ATCAAAAAATCCGCCGATAAATCCAGGGACTTTGTATTCCGCCTAG**TTAAAACTTCGG**

B

```
         10         20         30         40         50         60
MFIDVVPNGR SASAVLLRES FREGRKVHKR TIANLSQMPA ELVDGLRALL AGGSVVGGPD

         70         80         90        100        110        120
QALEIRRSLP HGHVAAVLGM MRKLEIPRLL GRQVSRERDL ALALIASRVI APGSKLSTLR

        130        140        150        160        170        180
GLNPETATSS LGQVLGLGVI EEREIYAALD WLGAQQGRIE RQFAKRHLRD GTLVLYDVSS
                                                              *
        190        200        210        220        230        240
SYLEGRCCEL AQHGYSRDHR PDRLQIVYGL LCDREGRPIA VEVFEGNTAD PGTIAAQVEK

        250        260        270        280        290        300
LKRRFHLNHV VLVGDRGMIT TARIRKEIKP AGLDWISCLR AGQIQDLAEG PLQMSLFDER
                 *
        310        320        330        340        350        360
DIAAIASPDY PGERLIACRN AALAGERRRK REALLTATER ELTRIVAATT RKRAPLRGAA

        370        380        390        400        410        420
EIGLAVGAVI NQRKMAKHFD LTITADRFSF RRNEAGIARE AALDGIYVIR TSVAAEAMSD

        430        440        450        460        470        480
ADTVRAYKDL SRVERAFRTL KSVDLAIRPV HHWLSPRVRA HVFLCMIAYY VEWHLRDALK
               *        *
        490        500        510        520        530        540
PILFQDHDPL AAEAERASPV APATISPAAK RKRGRRRNDD NLPLSSFADL MAHLATQTLN

        550        560        570
TAALPKAPNA TFTTLATPTT LQAAAFNLLE IEPMRVQ
```

Figure III.2. Codon usage of the IS*1634* gene.
Rare codons in *E.coli* are shadowed. 26 cases of codon CGG (Arg), 7 of CCC (Pro), 3 cases of CGA (Arg), 2 of GGA (Gly) and 1 of CTA (Leu). Translated amino acid sequence is also shown.

```
atgttcatcgacgtggttccgaatggccgctcggcgtcggcggtgctgctgcgggagagt
 M  F  I  D  V  V  P  N  G  R  S  A  S  A  V  L  L  R  E  S
ttccgtgaggggcggaaggtccacaagcgcacgatcgccaatctgagccagatgccggcg
 F  R  E  G  R  K  V  H  K  R  T  I  A  N  L  S  Q  M  P  A
gagctggtcgatggcttgcgcgccctgctcgccggcggctcggtggtgggcggcccggat
 E  L  V  D  G  L  R  A  L  L  A  G  G  S  V  V  G  G  P  D
caggcgctcgagatccggcgatccctgccgcacgggcacgtggcggcggtgctggggatg
 Q  A  L  E  I  R  R  S  L  P  H  G  H  V  A  A  V  L  G  M
atgcgcaagctggagattccgcgcctgctgggacgccaggtctcgcgcgagcgggacttg
 M  R  K  L  E  I  P  R  L  L  G  R  Q  V  S  R  E  R  D  L
gcactggcactgatcgcgagtcgcgtgatcgcgccgggctcgaaactctcgacgctgcgc
 A  L  A  L  I  A  S  R  V  I  A  P  G  S  K  L  S  T  L  R
ggcctgaacccggagacggcgacctcgagcctcgggcaggtgcttgggctcggcgtgatt
 G  L  N  P  E  T  A  T  S  S  L  G  Q  V  L  G  L  G  V  I
gaggagcgcgagatctacgccgccctcgactggttgggcgcgcagcaggggcggatcgaa
 E  E  R  E  I  Y  A  A  L  D  W  L  G  A  Q  Q  G  R  I  E
cggcagttcgcgaagcgccatctgcgcgatggcacgctggtgctctatgatgtcagctcg
 R  Q  F  A  K  R  H  L  R  D  G  T  L  V  L  Y  D  V  S  S
tcctatcttgaaggccggtgctgcgaactcgctcaacacggttatagccgtgatcatcgg
 S  Y  L  E  G  R  C  C  E  L  A  Q  H  G  Y  S  R  D  H  R
ccggatcggctccagatcgtctatggtctgctgtgcgatcgggagggccggccgatcgcg
 P  D  R  L  Q  I  V  Y  G  L  L  C  D  R  E  G  R  P  I  A
gtcgaggtgttcgaaggcaacaccgccgatcccggcacgatcgcggcccaggtggagaag
 V  E  V  F  E  G  N  T  A  D  P  G  T  I  A  A  Q  V  E  K
ctcaagcggcggtttcatctgaaccacgtcgtgctggtgggcgatcggggcatgatcacc
 L  K  R  R  F  H  L  N  H  V  V  L  V  G  D  R  G  M  I  T
acggcgcggatccgcaaggagatcaaacccgccggattggactggatcagttgcctgcgg
 T  A  R  I  R  K  E  I  K  P  A  G  L  D  W  I  S  C  L  R
gcgggtcagatccaggacctcgccgaggggccgctgcagatgtcgctgttcgacgagcgc
 A  G  Q  I  Q  D  L  A  E  G  P  L  Q  M  S  L  F  D  E  R
gatatcgccgcgatcgcatcgcctgattatcccggcgagcggctgatcgcctgccgcaat
 D  I  A  A  I  A  S  P  D  Y  P  G  E  R  L  I  A  C  R  N
gcggctctggccggggaacgccggcgcaagcgcgaggcgctgcttaccgcgaccgagcgg
 A  A  L  A  G  E  R  R  R  K  R  E  A  L  L  T  A  T  E  R
gaattgacccgcatcgtggcggcgacgacgcgcaagcgcgcgccgttgcgcggcgcggcc
 E  L  T  R  I  V  A  A  T  T  R  K  R  A  P  L  R  G  A  A
gagatcggccttgcggtcggtgccgtgatcaaccagcgcaagatggccaagcatttcgat
 E  I  G  L  A  V  G  A  V  I  N  Q  R  K  M  A  K  H  F  D
ctcaccatcaccgccgatcgcttcagcttccggcgcaacgaagccggtatcgcccgcgag
 L  T  I  T  A  D  R  F  S  F  R  R  N  E  A  G  I  A  R  E
gcggcgctcgatggcatttacgtcatccggaccagcgtcgccgcggaggcgatgagcgat
 A  A  L  D  G  I  Y  V  I  R  T  S  V  A  A  E  A  M  S  D
gccgacaccgtgcgggcctacaaggacctctcccgggtggaacgggcgttccgaaccctg
 A  D  T  V  R  A  Y  K  D  L  S  R  V  E  R  A  F  R  T  L
aaatcggtcgacctcgcaatccgtccggtccatcactggctctcgccgcgggtgcgcgcg
 K  S  V  D  L  A  I  R  P  V  H  H  W  L  S  P  R  V  R  A
catgtctttctctgcatgattgcctattacgtcgagtggcatctgcgcgatgccctcaag
 H  V  F  L  C  M  I  A  Y  Y  V  E  W  H  L  R  D  A  L  K
ccgatcctgtttcaggatcacgatccactggccgccgaggccgagcgcgcctccccggtt
 P  I  L  F  Q  D  H  D  P  L  A  A  E  A  E  R  A  S  P  V
gcccctgccacgatctcacccgccgccaagcgcaagcgggggcgacgccgcaacgacgac
 A  P  A  T  I  S  P  A  A  K  R  K  R  G  R  R  R  N  D  D
aacctgccgctctcaagctttgccgacctgatggcgcatctggccacccaaactctcaac
 N  L  P  L  S  S  F  A  D  L  M  A  H  L  A  T  Q  T  L  N
accgccgcgctgcccaaggcgcccaatgcgaccttcaccaccctggccacgccaaccaca
 T  A  A  L  P  K  A  P  N  A  T  F  T  T  L  A  T  P  T  T
ctacaggcggccgccttcaacctcctcgaaatcgaacccatgcgtgtccagtag
 L  Q  A  A  A  F  N  L  L  E  I  E  P  M  R  V  Q  -
```